Reimagining the Central Challenge of Face Recognition: Turning a Problem Into an Advantage

Ognjen Arandjelović [a]

School of Computer Science

University of St Andrews

St Andrews KY16 9SX

United Kingdom

[a] *Tel: +44(0)1334 46 28 24*

E-mail: `ognjen.arandjelovic@gmail.com`

Web: `http://oa7.host.cs.st-andrews.ac.uk/`

Abstract

High inter-personal similarity has been universally acknowledged as the principal challenge of automatic face recognition since the earliest days of research in this area. The challenge is particularly prominent when images or videos are acquired in largely unconstrained conditions 'in the wild', and intra-personal variability due to illumination, pose, occlusions, and a variety of other confounds is extreme. Counter to the general consensus and intuition, in this paper I demonstrate that in some contexts, high inter-personal similarity can be used to advantage, i.e. it can help *improve* recognition performance. I start by a theoretical introduction of this key conceptual novelty which I term 'quasi-transitive similarity', describe an approach that implements it in practice, and demonstrate its effectiveness empirically. The results on a most challenging real-world data set show impressive performance, and open avenues to future research on different technical approaches which make use of this novel idea.

Key words: Meta-algorithm, paradigm change, retrieval, intra-class, inter-class, similarity, dissimilarity.

1 Introduction

Face recognition is often described as one of the most active areas of research in computer vision [1, 2, 3, 4]. While I am unaware of attempts to formalize this claim and support it with rigorous empirical evidence, it is beyond doubt that the field has undergone substantial changes over time. By this I am not referring merely to changes in the technical approach which can be naturally expected to take place as advances are made, but rather to the practical paradigms and the context in which face recognition is employed.

Early face recognition work can be described as a proverbial exploratory mission which served to deepen the understanding of the key challenges and features (in an abstract sense) which have the greatest discriminative power [5, 6]. Geometric features and the first statistical appearance based methods were described in this period. Thereafter the focus has shifted to the practical challenge of making face recognition useful in real world security oriented applications. It is in this period that the difficulty of the problem has crystallized, with concurrent changes in pose, illumination, resolution, and other extrinsic factors, exposing the limitations of the proposed algorithms [7, 8, 9, 10]. Most face recognition work falls under the umbrella of this conceptual period. Despite the immense amount of research effort, both by academia and industry, the highly optimistic predictions expressed in the early years of face recognition research failed to materialize: in unconstrained conditions the performance of face recognition in security applications remains disappointing [11, 12, 13]. The key reason lies in the nature of the demands of most security applications on the one hand, and the inherent discriminative weakness of facial biometrics. As regards the former, security applications demand a low false positive rate (allowing an intruder the access to a resource carries a high cost) and often a low false negative rate (denying access to a legitimate user is frustrating, time consuming, and potentially costly). At the same time, on the latter point, there is no compelling evidence that face based biometrics even in principle can be used to attain these demands. Face recognition by humans, often intuitively seen as highly sophisticated, is in fact not very accurate when evaluated in conditions comparable to those in which automatic methods are expected to operate [14, 15]. Humans use a variety of constraints, such as knowledge based priors ('whom do I expect to encounter in this place?'), complementary biometrics (height, gait, voice, etc.), and a plethora of others to simplify the task in everyday situations. However, such assumptions are either difficult to incorporate in automatic methods (e.g. due to the semantic gap) or inappropriate in the context of practical applications of interest. While work on the underlying fundamentals continues with unabated effort [16, 17, 18, 19, 20], with particularly promising innovations arising from the use of sparse coding [2, 19, 21], dictionary representations [22, 23], and deep learning [24, 25, 26, 27], turning point for face recognition research has come in the last decade with the emergence of massive amounts of visual data – the focus has shifted to the use of face recognition for the retrieval and organization of photographs and video recordings [28, 4, 27]. The requirements of these applications contrast the aforementioned requirements of security applications: following the successes of web search engines, by adopting the ranked retrieval presentation of output, both so-called type I and type II errors are much more readily tolerated. The user is often not overly troubled by not every instance of interest being retrieved, or it not being retrieved at rank-1, as long as correct matches are within a reasonable rank (the quantified meaning of 'reasonable' being somewhat dependent on the application).

Thus, to summarise briefly the history of face recognition, the field has largely been characterized by incremental (but important and cumulatively significant) technical advances with major practical leaps which came though by innovative ways of seeing the same problem though a different lens. In the present paper my aim is to achieve the latter. Specifically, I will argue from theory that a characteristic at the heart of all face recognition problems, which is universally considered as *the* key challenge, can in fact be turned into an advantage in the right context. My case is first put forward on rigorous theoretical grounds, and subsequently demonstrated and discussed using empirical evidence.

The broad topic of the present paper is that of face set retrieval and the central contribution relates both to the previous work on set based recognition and the work concerned with recognition in the context of large data collections [28, 4, 27]. In contrast to most work in the literature herein my principal interest is neither in the representation of face sets nor in the associated similarity measures *per se*. Rather, given a baseline algorithm for measuring the similarity of two face sets, I seek to leverage the structure of the data at a large scale, that of the entire database, to make the best use of the available baseline. In the sense that the proposed method has as its input both data (face image sets) and an algorithm (the 'baseline'), it can be accurately thought of as a *meta-algorithm*.

1.1 Problem statement

Given a query face set the aim is to retrieve image sets of the same person from a large database (the 'gallery'). More specifically, the desire is to order the gallery sets in decreasing order of confidence that they match the query by identity. Thus the ideal retrieval has all sets of the query person first ('matches') followed by all others ('non-matches'). I assume that the gallery is entirely unlabelled and may contain multiple sets of the same person.

2 Learnt transitive similarity

In this section I introduce the main contribution of the paper. In particular, I describe a general framework for face retrieval especially well suited for large collections of face images acquired 'in the wild' i.e. in largely unconstrained imaging conditions, and characterized by highly unbalanced amounts of training data per class (person). I start by motivating the intuition behind the proposed method in the section which follows, and subsequently explain how this intuition can be formalized into a general retrieval framework.

2.1 Motivation and the key idea

It is insightful to begin by considering the motivation behind the key idea in the context of related previous work and in particular the Matched Background Similarity (MBS) method of Wolf *et al.* [29]. In brief, Wolf *et al.* argue that in building a classifier which discriminates the appearance of a specific person from that of all other people, the focus should be on discriminating between this person and those individuals most similar to them; improvements in discrimination against very dissimilar people matter less as these individuals are unlikely to be conflated with the person of interest anyway. The idea I introduce here can be seen as complementary and builds upon a similarly simple basic principle. Specifically, I make use of the observation that if person A is alike in appearance to person B, and similarly person B to person C, *on average* persons A and C are more likely to look alike than two randomly chosen individuals. I term this Quasi-Transitive Similarity, the prefix 'quasi-' capturing the notion that the stated regularity is a statistical rather than a universal one, as I shall explain shortly.

This is illustrated conceptually in Figure 1 using images of the former prime minister of Australia, Tony Abbot, and the actor Daniel Craig. For clarity, the variability of a person's appearance is shown as a 1D manifold. Specifically, the manifolds shown in black represent the appearance variability within the corresponding sets. The dotted manifold shown in red represents the range of appearance of Tony Abbott which is present neither in the gallery nor in the query set (in this conceptual example these are left semi-profile to left profile images).

As stated in the introduction above, the transitivity of similarity in appearance does not hold universally. It is possible that persons A and B are similar by virtue of one set of physical

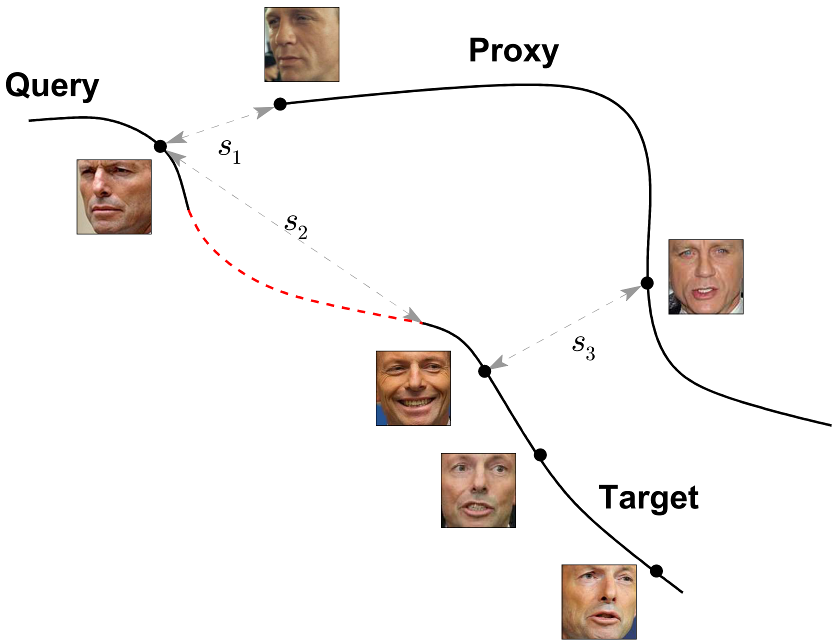


Fig. 1. The similarity between a query and the correct target set (initially poorly matched) may be better estimated indirectly via proxy data. 1D manifolds shown in black represent the appearance variability within sets. The dotted manifold shown in red represents the range of appearance of T Abbott present neither in the gallery nor in the query set. The query is poorly matched to the correct set because the person's pose in the query is vastly different than any of the poses in the target set. However, the query matches well the proxy set which contains more extensive pose variability of a person similar in appearance to the target person, the said similarity being directly inferable from data from the similarity of the matched images in the two sets.

features, and B and C by virtue of another. A useful mental picture can be formed by drawing an analogy from statistics (or geometry): random variables (or vectors) A and B, and B and C may be positively correlated (have a positive dot product), yet A and C may be negatively correlated (have a negative dot product) with one another. This is illustrated in Figure 2.

Lastly, it is worth contrasting the present approach with that of Yin *et al.* [30]. Unlike the method herein, their method necessitates the localization of face parts, which is problematic and highly likely to fail in severe illuminations, extreme poses, or in poor quality images.

Their method also needs to extract estimates of pose and illumination, again very much unlike the one proposed herein which does not have any of the aforementioned bottlenecks – all learning is performed directly from data and without the need for an explicit model at a higher semantic level.

2.2 *Transitivity meta-features*

I have already noted that the observed transitivity of similarity is a statistical rather than a universal phenomenon. In other words, while the similarity of persons A and B, and B and C, *on average* leads to a greater similarity between A and C, in some instances this will not be the case. This observation suggests that in addition to inter-personal similarities between persons A and B, and B and C, a richer set of features should be used to *infer* the similarity between persons A and C. By implication, these features should complement the inter-personal similarities in the sense that jointly they should allow for a better estimate of the similarity between persons A and C than just similarities between persons A and B, and B and C, or indeed the direct baseline comparison of persons A and C (i.e. without the use of additional indirect information provided by the relationship of B with A and C).

To motivate the meta-features that I propose in the present work, consider the conceptual illustrations shown in Figure 3. Solid coloured lines depict the range of appearance variation within face sets. The aim is to estimate the similarity of the query (green) and the set denoted as ‘target’ (red). To clarify, by a ‘target’ set we mean any gallery set which as such may be a potential correct match. The face set marked ‘proxy’ is a database face set of a person similar in appearance to the ‘target’, as assessed by the baseline similarity measure; for example, the proxies of a particular target set can be selected as its nearest k_p sets in the database. The dotted red line represents the range of possible appearance of the ‘target’ person which is not actually present in the ‘target’ face set. For the time being the reader may assume that face

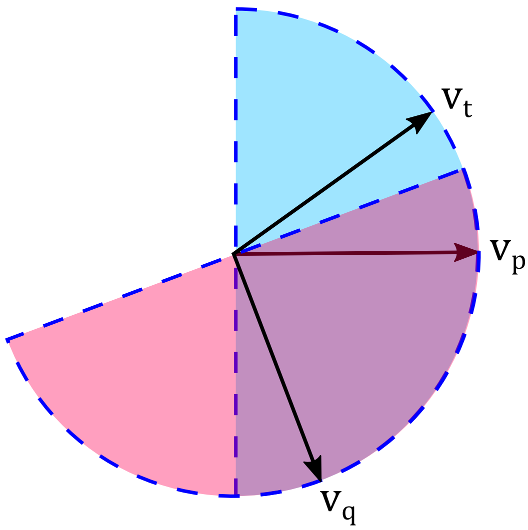
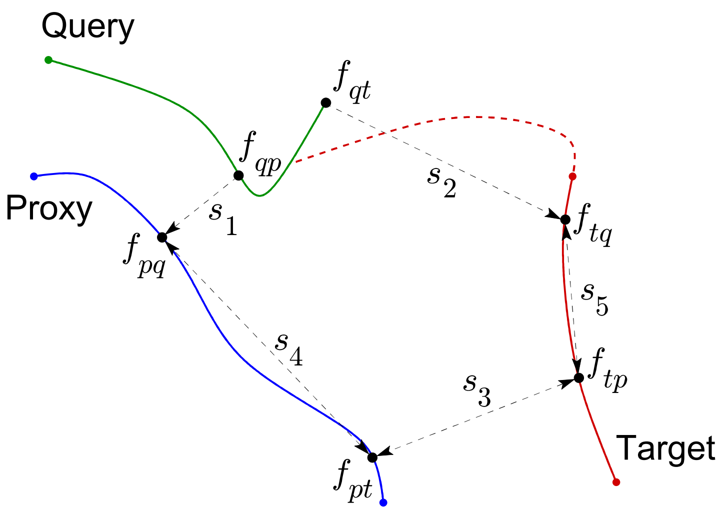


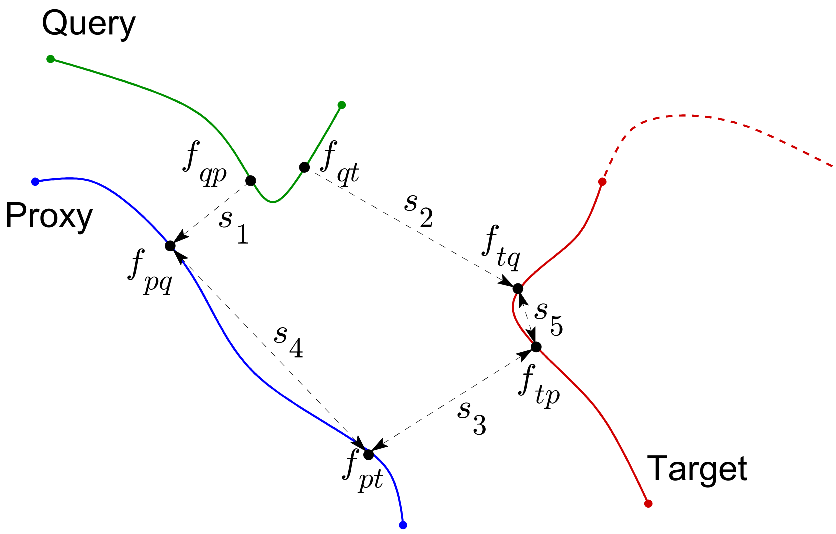
Fig. 2. A conceptual illustration of the non-universality of transitivity of pair-wise similarity. Shown are three vectors in two dimensions: v_q, v_t, and v_p. The red and blue shaded semicircles indicate the angle ranges within which vectors have a positive dot product with respectively v_q and v_p. Observe that although the dot product between v_q and v_p is positive (i.e. the two vectors can be regarded as exhibiting a degree of similarity), as is the dot product between v_p and v_t, the dot product between v_q and v_t is negative.

sets are represented as sets of actual exemplars and the similarity between two sets is given by the similarity between their most similar members – I will explain how the ideas introduced herein can be generalized in the next section.

Both in the case shown in Figure 3(a) and that in Figure 3(b), the baseline similarity measure tells us that 'query' is close to 'proxy', and of course 'proxy' is close to 'target' by design i.e. by the former being a proxy in the first place. The difference between the two cases, illustrated conceptually, lies in the similarity of exemplars f_{tq} and f_{tp} i.e. the exemplars best matching the query and proxy sets. In particular, the observation that the baseline similarity measure deems the proxy set significantly more similar than the query to the target on the



(a) Query and target: same identity



(b) Query and target: different identities

Fig. 3. Transitivity meta-features extracted using a baseline set comparison: conceptual motivation, using (a) a matching (same identity) query-target set pair, and (b) a non-matching (differing identities) query-target set pair.

one hand, while both similarities are explained by similar target exemplars, informs us that the divergence in query and proxy appearances from the target are of different natures. Thus, even if similarities s_1, s_2, and s_3 are the same in Figure 3(a) and Figure 3(b), the information contained in relationships between f_{tq} and f_{tp}, and f_{pq} and f_{pt} tells us that we should infer different query-target similarities in the two cases. Therefore I introduce what I term transitivity meta-features which I use for the aforementioned inference. Given a baseline similarity measure and a triplet consisting of query, target, and proxy sets, the corresponding transitivity meta-feature $v(\text{query,target}|\text{proxy})$ comprises five similarities – s_1 ('query' to 'proxy' similarity), s_2 ('query' to 'target' similarity), s_3 ('proxy' to 'target' similarity), s_4 (similarity between the 'proxy' exemplar most similar to 'query' and the 'proxy' exemplar most similar to 'target'), and s_5 (similarity between the 'target' exemplar most similar to 'query' and the 'target' exemplar most similar to 'proxy'):

$$v(\text{query,target}|\text{proxy}) = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{bmatrix} \quad (1)$$

2.3 *Non-exemplar based representations*

In the preceding discussion I asked the reader to think of appearance variation within each set as being represented using what is probably conceptually the simplest choice of representation: as a collection of exemplars. In other words, each set was a set of representations of individual faces. This was done for pedagogical reasons and I now show that the proposed framework is in no way reliant on this representation.

In particular, to make the transition of applying the proposed method on the special case in which a face set is represented using a set of directly observed exemplars to the general case in which an arbitrary set representation is employed, I need to explain how the concept of a pair of the most similar exemplars such as those labelled f_{qp} and f_{pq} in Figure 3(a), as well as the similarity between them (such as that between f_{pq} and f_{pt}), can be generalized. This is not difficult – all that is required is a slight reframing of the concept. Instead of seeking the nearest pair of specific exemplars, in the general case we are interested in the pair of the most similar modes of variation captured by the representations of two sets (as measured by the baseline similarity measure of course). I illustrate this idea with a few examples.

If the variation within a set is modelled using a linear subspace and the subspace-to-subspace generalization of the distance from feature space (DFFS) [31] adopted as the (dis)similarity measure between them, the most similar modes of variation between two sets represented using such subspaces are sub-subspaces themselves [**?**]. These correspond to different exemplars f_{xy} in Figure 3 and can be compared using the DFFS baseline. If, on the other hand, similarity is measured using the maximum correlation between subspace spans [32], the most similar modes of variation between two sets are readily extracted as the first pair of the canonical vectors between subspaces [33] and compared using the cosine similarity measure [34, 35]. For manifold-to-manifold distances such as that of Lee *et al.* [36] the most similar modes of variation are simply the nearest pairs of points on two manifolds, with the similarity of two points on the same manifold readily quantified by the geodesic distance between them.

The same ideas are readily applied to any of a variety of set representations and similarity measures described in the literature.

2.4 Learning quasi-transitive similarity

Given a triplet comprising a query, a target, and a proxy data set, our aim now is to infer the similarity between the query and the target using the corresponding transitivity feature defined in (1). Without loss of generality, let us quantify inter-set similarity with a real number in the range $[0, 1]$, where 0 signifies the least and 1 the greatest possible similarity. Then the problem can be stated formally by saying that we are seeking a mapping m_{qts}:

$$m_{\text{qts}} : \mathbb{R}^5 \to [0, 1], \tag{2}$$

with the ideal output of $m_{\text{qts}}(v(\text{query,target}|\text{proxy}))$ being 0 iff the identities in the query and target sets are different, and 1 iff they are the same. Observe that since we are interested in confidence based ranking of all sets in a database, the codomain of m_{qts} is not the set $\{0, 1\}$, which would make this a binary classification problem, but rather $[0, 1]$ (a range) which makes it a regression task.

In the types of problem setting in which face recognition is addressed by most of the existing research, obtaining features for training, at least in principle, is simple. Whether it is verification (1-to-1 matching) or identification (1-to-N matching), the database 'known' to the algorithm comprises data which is, it is assumed, correctly partitioned by the identity. The retrieval setting adopted in this work is more challenging in this sense and consequently the learning process needs to be approached with more care. In particular, as described in Section 1, I assume that the database is entirely unlabelled and that it may contain multiple sets of the same person. We neither know how many individuals there are in the database nor the number of sets of each individual (which can of course vary person to person). Since for any two database sets we cannot know for certain if they belong to the same or different individuals, an obvious corollary is that in the extraction of transitivity meta-features described by (1) both intra-personal and inter-personal training sets may contain incorrect examples.

2.4.1 Extraction of transitivity meta-features for training

Given that our data is unlabelled i.e. that we do not know if two face sets in the database correspond to the same person or not, we cannot extract training transitivity meta-features in the obvious manner by considering different query, target, and proxy triplets, with the query and the target either matching (producing same identity training data) or not (producing differing identities training data). Instead, I describe how training data, albeit corrupted (this issue is dealt with in the next section), can be collected automatically by considering only pairs of sets, that is, all possible database sets and their proxies. I do this for the two baseline set comparison methods adopted from the work by Wolf *et al.* [29] (described in more detail in Section 3.3):

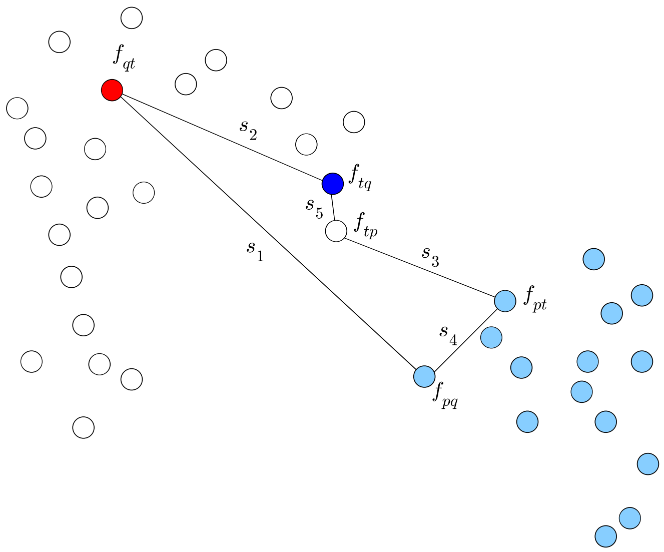
- The *maximum maximorum* cosine similarity between sets of exemplars [37], and
- The maximum correlation between vectors confined to linear subspaces describing within set variability [38].

For the benefit of the reader and as an additional illustration of the generalizability of the approach, automatic training data extraction for use with the Extended Canonical Correlation Analysis (E-CCA) based baseline is included in Appendix A.

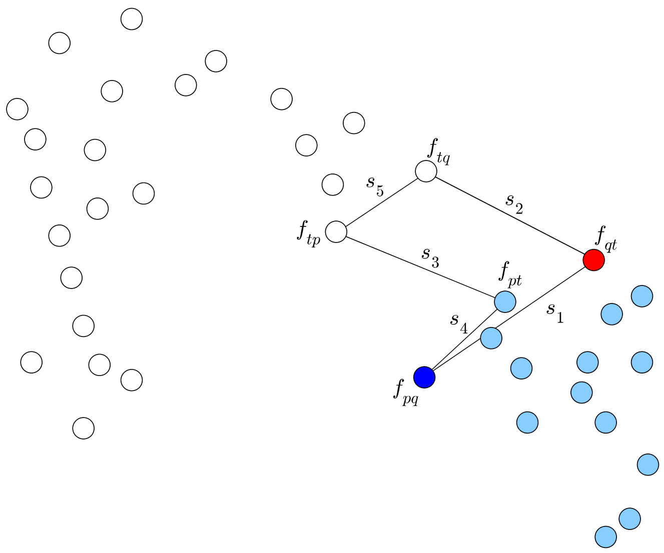
Exemplar based baseline Consider a particular database face set ('reference') used for training and one of its proxies. To extract training transitivity meta-features which correspond to same identity query-target comparisons, I select *both* query and target data from the reference set (i.e. a single video). In particular, I treat all possible pairs of exemplars in the reference set as possible pairs f_{qt} and f_{tq}. Indeed, for specific choices of possible query and reference sets, any two appearances may present themselves as the nearest exemplars in them. The second element s_2 in the transitivity meta-feature is then simply given by the similarity between the two exemplars. On the other hand the similarity s_1 between the query and the proxy is given by the similarity between the unitary set consisting of the reference set exemplar treated as f_{qt} and the proxy set. The nearest proxy exemplar to f_{qt} is of course f_{pq}. The similarity s_3 is simply computed as the similarity between the reference set and the proxy, which also gives us exemplars f_{pt} and f_{tp}, and allows for a straightforward computation of s_4 (as the similarity between f_{pq} and f_{pt}) and s_5 (as the similarity between f_{tq} and f_{tp}). A single pair of reference and proxy sets thus gives us $n_r(n_r - 1)$ 'positive' training transitivity meta-features, where n_r is the number of faces in the reference set. The process is illustrated conceptually in Figure 4(a).

The extraction of training transitivity meta-features which correspond to differing identities query-target comparisons is similar. Now I iterate through all exemplar pairs of the proxy set, taking each pair as f_{qt} and f_{pq} in turn. The closest target exemplar to f_{qt} becomes f_{tq}, while f_{pt} and f_{tp} are determined as before, allowing for all transitivity meta-feature entries (exemplar similarities) to be computed as in the case of same identity query-target training data extraction. A single pair of reference and proxy sets thus gives us $n_p(n_p - 1)$ 'negative' training transitivity meta-features, where n_p is the number of faces in the proxy set. The process is illustrated conceptually in Figure 4(b).

It is important to observe that the set of 'negative' training transitivity meta-features extracted in the described manner may be corrupt. This is an inherent consequence of the problem setting – since the database is entirely unlabelled we cannot know if the identities of the people in the reference and proxy set are actually different. The proposed process of training the regressor, described in Section 2.4.2, takes this into account. Nevertheless, the amount of improvement achieved with the proposed method over its baseline is tied to the proportion of 'negative' training data which is incorrect – the improvement inevitably decreases as this proportion is increased. However, if this is so, i.e. if a great proportion of proxies of sets in the database actually represent the same identity as the sets they are proxies to, this by design means that the baseline comparison is very good to start with so no significant improvement can be reasonably expected. Thus, the proposed method is particularly attractive in challenging conditions in which the baseline classifier does not perform well.



(a) Exemplar based matching: obtaining positive training samples



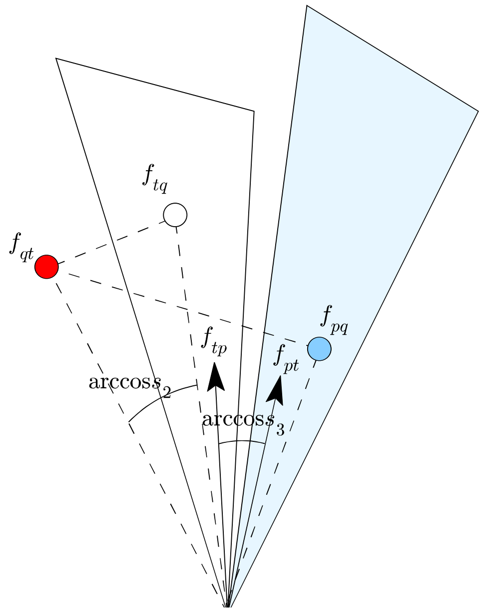
(b) Exemplar based matching: obtaining negative training samples

Fig. 4. Conceptual illustration of the proposed methodology for automatic collection of training data for training the quasi-transitivity regressor using the exemplar based baseline, from an unlabelled corpus. White and light blue data points respectively represent exemplars from a single face data set and one of its proxy sets. The red and dark blue points are randomly chosen images in an iteration of the algorithm (see main text for detailed explanation).

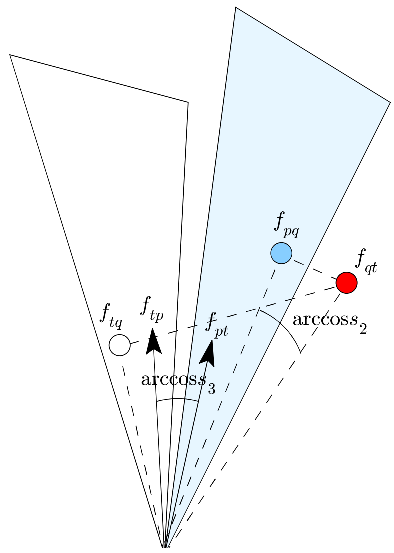
Subspace based maximum correlation baseline The extraction of training data for this representation is somewhat simpler than in the previous case. I again extract transitivity meta-feature training data using only face set pairs (rather than triplets) which are now represented by linear subspaces. To extract training transitivity meta-features which correspond to same identity query-target comparisons, I iterate through all reference set exemplars as f_{qt} and obtain f_{tq} and f_{pq} by projecting them to respectively the reference and proxy subspaces. Vectors f_{pt} and f_{tp} are readily obtained using the baseline set comparison as the principal vectors of the subspaces corresponding to reference and proxy subspaces. A single pair of reference and proxy sets thus gives us n_r 'positive' training transitivity meta-features. The process is illustrated conceptually in Figure 5(a).

The extraction of training transitivity meta-features which correspond to differing identities query-target comparisons proceeds in exactly the same manner, with the difference that it is proxy set exemplars that are iterated through as f_{qt} (as before also taken to be f_{qp}). A single pair of reference and proxy sets gives us n_r 'positive' training transitivity meta-features, where n_r is the number of faces in the reference set, and n_p 'negative' training transitivity meta-features, where n_p is the number of faces in the proxy set. A single pair of reference and proxy sets thus gives us n_p 'negative' training transitivity meta-features. The same remarks as before regarding the corruption of the 'negative' training set hold here too. The process is illustrated conceptually in Figure 5(b).

Closing notes and observations In Section 2.1 I remarked that the basic idea behind the proposed method can be seen as complementary to those of Wolf *et al.* [29]. However, when the proposed training scheme is considered it can be seen to contain both conceptually similar elements *and* complementary elements to MBS. In particular, since the negative training set of quasi-transitivity meta-features is extracted by considering elements of the proxy set as the query, the proposed method learns to discriminate precisely between a person and those



(a) Subspace alignment based matching: obtaining positive training samples



(b) Subspace alignment based matching: obtaining negative training samples

Fig. 5. Conceptual illustration of the proposed methodology for automatic collection of training data for training the quasi-transitivity regressor using the baseline based on the maximum correlation between subsets, from an unlabelled corpus. The white and light blue subspaces respectively correspond to a single face data set and one of its proxy sets. The red point is a randomly chosen image in an iteration of the algorithm (see main text for detailed explanation).

individuals most similar to him/her (as in MBS), while exploiting the quasi-transitivity of similarity (complementary to MBS).

2.4.2 Quasi-similarity predictor design

In this paper I propose the use of the ϵ support vector (ϵ-SV) regression [39]. For comprehensive detail of this regression technique the reader is referred to the original work by Vapnik (also see Schölkopf and Smola [40]); for the sake of completeness and continuity, herein I present a brief summary of the ideas relevant to the proposed method.

Given training data $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset \mathcal{F} \times \mathbb{R}$, where $\mathcal{F}$ is the input space (in our case this is $\mathbb{R}^5$), ϵ-SVR aims to find a function $h(x)$ which deviates at most ϵ from its targets y. As in other SV based methods, an implicit mapping of input data $x \to \Phi(x)$ is performed by employing a Mercer-admissible kernel [41] $k(x_i, x_j)$ which allows for the dot products between mapped data to be computed in the input space: $\Phi(x_i) \cdot \Phi(x_j) = k(x_i, x_j)$. The function $h(x)$ of the form

$$h(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{3}$$

is then learnt by minimizing

$$\sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j) \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) - \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \tag{4}$$

subject to the constraints $\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0$ and $\alpha_i, \alpha_i^* \in [0, c]$. The parameter c can be seen as penalizing prediction errors greater than ϵ i.e. as balancing the trade-off between the smoothness of $h(x)$ and the amount of data predicted with an error greater than ϵ.

The nature of ϵ-SV regression is particularly well suited to the problem at hand. The key insight stems from the observation that since we are not looking to make a crisp decision on whether people's identities are the same, but rather derive a confidence measure thereof.

Hence, I train the regressor using the value of 1 as the target for same identity transitivity meta-features, and 0 for different identities, allowing for a large prediction error margin of $\epsilon = 0.4$ but severely penalizing greater errors by setting $c = 1000$. The large penalty c ensures that it is the outliers in the form of the wrongly labelled training data that define the boundary between the penalized and non-penalized regions of the high-dimensional space, while the wide margin $\epsilon = 0.4$ ensures that the correctly labelled bulk of the training corpus is pushed away from the boundary towards the desired extreme values of 0 and 1. I used the radial basis function kernel $k(x_i, x_j) = \exp\{-0.2\|x_i - x_j\|^2\}$.

A schematic illustration of the overall learning of quasi-transitivity, underlay by a specific adopted baseline set based comparison, is shown in Figure 6.

2.4.3 Retrieval

Given a query data set I compute its similarity with a target database set by computing the regression based estimate $m_{\text{qts}}(v(\text{query,target}|\text{proxy}))$ using each of target's k_p proxies, and taking the maximum of these and the baseline similarity between the query and the target. Database sets are then ordered by decreasing similarity with respect to the query. This is schematically illustrated by the diagram in Figure 7

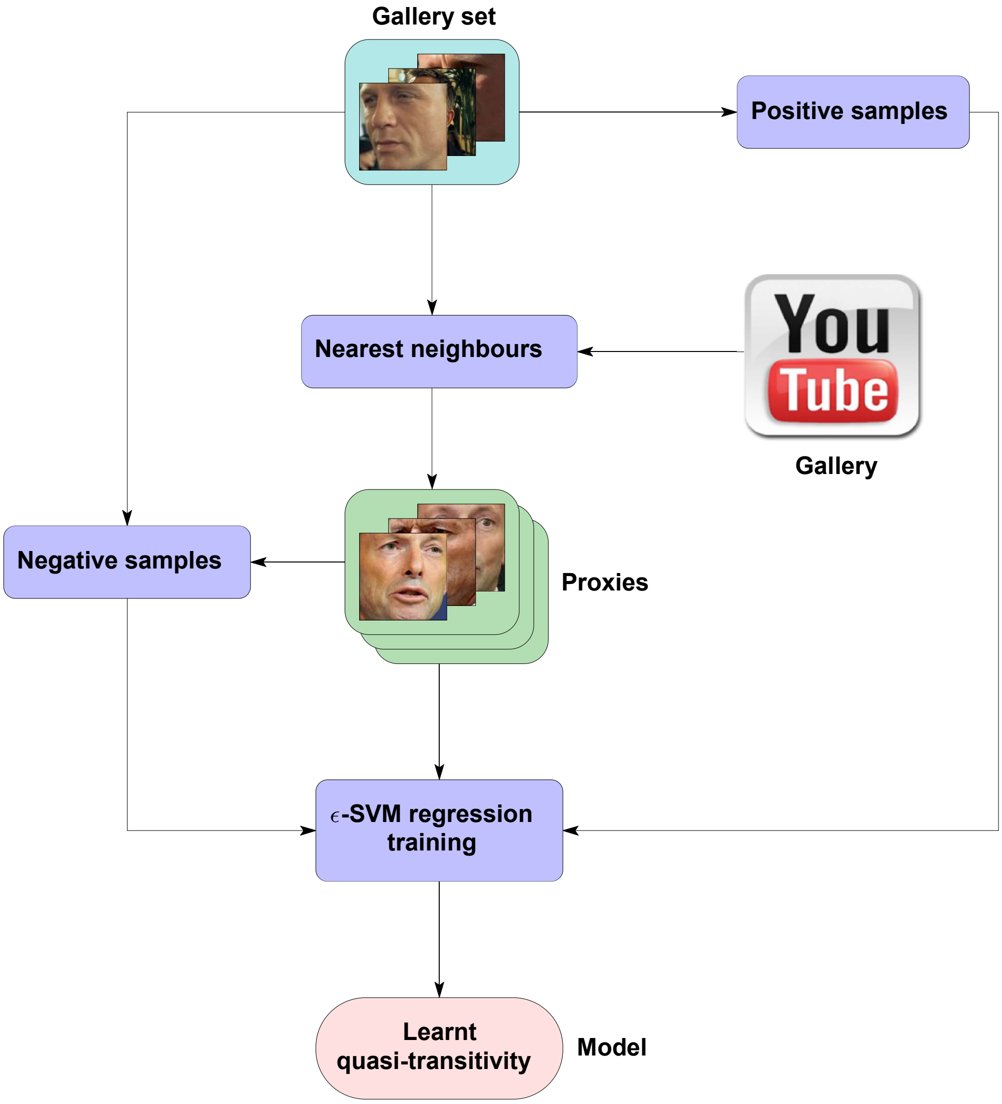


Fig. 6. Schematic overview of the proposed meta-algorithm training stage. For each set in the gallery (the 'target' set), a set of proxies is automatically extracted first; see Section 2.2 for comprehensive detail. Then, negative (different identity) meta-features (see Section 2.2) samples are extracted from the target set and its proxies, as positive samples from the target set alone, which is a process dependent on the adopted baseline set based comparison algorithm; detailed examples are given in Section 2.4.1. These used to learn the introduced quasi-transitivity i.e. the meta-algorithmic model.

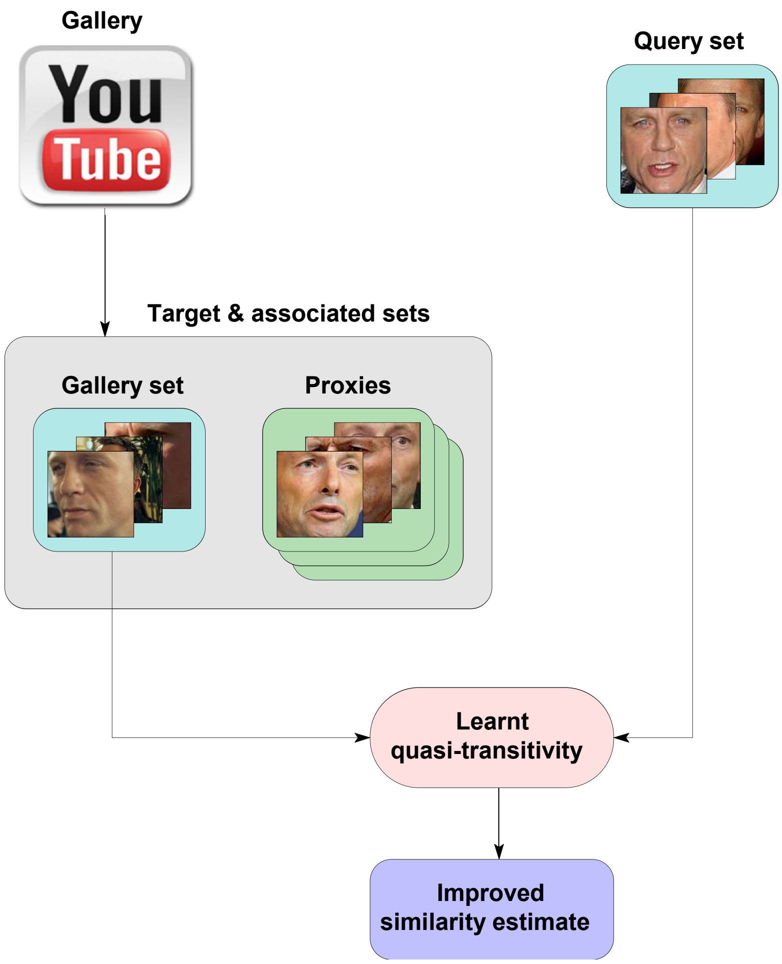


Fig. 7. Schematic overview of the proposed meta-algorithm querying (application) stage. For each set in the gallery (the 'target' set), the learnt meta-algorithmic model (also see Figure 6) is applied to compute an improved similarity estimate, using the adopted baseline set based comparison.

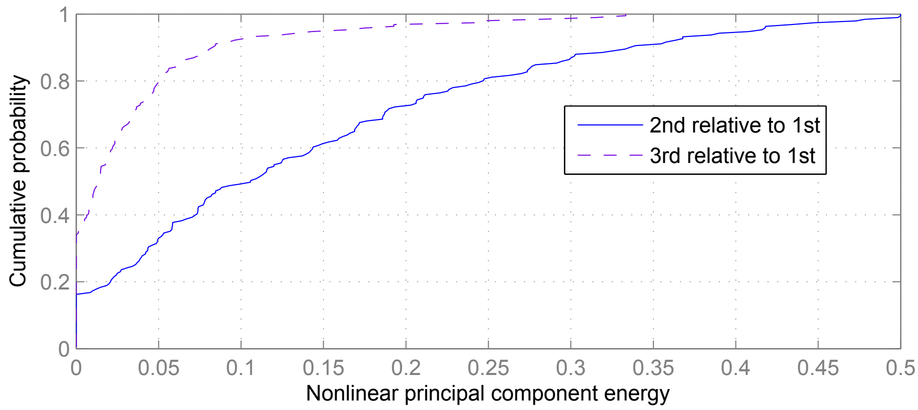


Fig. 8. The cumulative distribution function (CDF) of the data energy contained in the 2nd and 3rd nonlinear kernel PCA components relative to the energy of the 1st component, across sets in the YouTube Faces Database. The variation within sets is strongly dominated by the 1st nonlinear principal component.

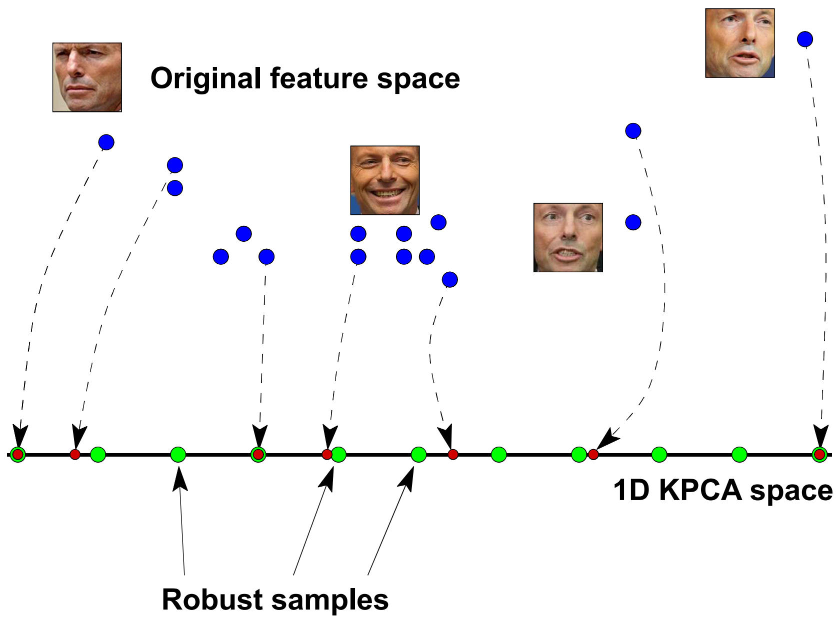


Fig. 9. Conceptual illustration of the proposed robust sample selection: (i) original exemplars are projected onto their 1st kernel principal component, (ii) uniform sampling between the extreme projections is performed in the 1D kernel space, and (iii) the obtained samples are re-projected into the original space (step not shown).

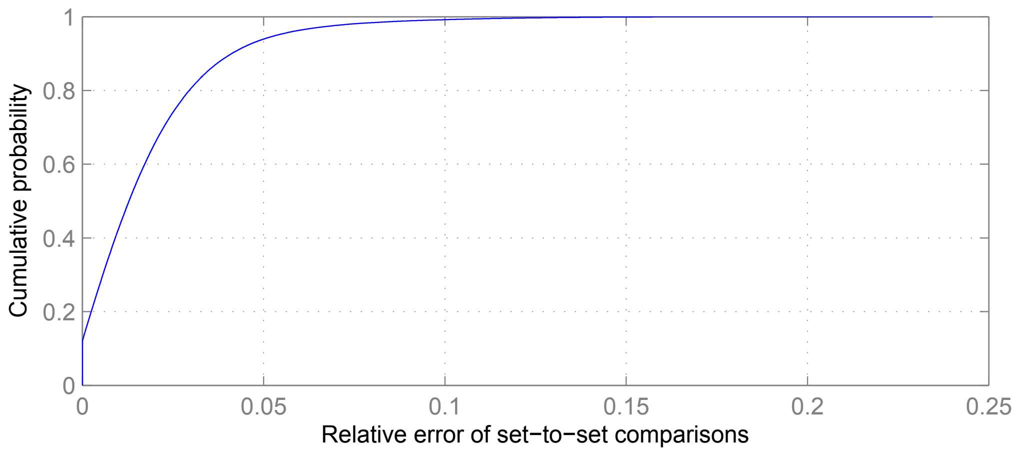


Fig. 10. CDF of the error introduced by the proposed robust sample selection (10 samples were used) in the exemplar based set method. Also see Figure 9.

3 Evaluation

In this section I report my evaluation of the proposed methods and discuss my findings. I start by describing the data set on which the evaluation was performed, consider the measures used to assess performance, summarize the evaluated baseline set representations and distances, and finally present and discuss the results.

3.1 Evaluation data

For evaluation I adopted the YouTube Faces Database [29] which contains sets of faces extracted from YouTube videos. There are two key reasons which motivated this choice. Firstly, the manner in which this data set was collected and the nature of its contents are representative of the conditions which the present work targets. In particular, the total amount of data is large (3425 face image sets of 1595 individuals, with the average set size of approximately 181.3 faces or equivalently 620,953 faces in total), it was extracted from videos acquired in unconstrained conditions in which large changes in illumination, pose, and facial expressions are present, and the distribution of data is heterogeneous both with respect to the set sizes (48–6,070) as well as the number of sets (1–6) for each person in the database. The second reason lies in the reproducibility of results and the ease of comparison with alternatives in the literature – the database has been widely adopted as a standard benchmark and a number of standard face representations are provided ready for use. Full detail can be found in the original publication [29].

3.2 *Performance evaluation*

As the cornerstone measure of retrieval performance I adopt the widely used average normalized rank (ANR) [42, 43, 44]. In brief, ANR treats each retrieved datum as either matching or not matching the query and computes the average rank of the former group, normalized to the range $[0, 1]$, with the ANR value of 0 corresponding to the best possible performance (all matching data retrieved before any non-matching) and 1 the worst (all non-matching data retrieved before any matching). Formally:

$$ANR(n, \{r_1, \ldots, r_c\}) = \frac{\sum_{i=1}^{c} r_i - m}{M - m} \tag{5}$$

where n is the database size, $\{r_1, \ldots, r_c\}$ the set of retrieval ranks corresponding to the data of interest (i.e. data matching the query), and m and M respectively the minimum and maximum possible values of the sum of $r_1, \ldots, r_c$:

$$m = \sum_{i=1}^{c} i = \frac{c \times (c+1)}{2} \tag{6}$$

$$M = \sum_{i=n+1-c}^{n} i = c \times \frac{2n - c + 1}{2} \tag{7}$$

In comparison with other common performance measures, such as the receiver operating characteristic (ROC) curve [45, 46], commonly used in verification and identification problems (including Wolf *et al.* [29]), the average normalized rank more directly captures the ultimate aim of a retrieval algorithm. While a detailed discussion of this topic is outside of the scope of the present paper, note additionally that ANR reflects retrieval performance *better* too – it is possible, for example, for all possible retrievals on a data set to be best possible (correct matches always retrieved first) with the ROC curve exhibiting non-ideal behaviour.

3.3 Methods

Motivated by the results reported by Wolf *et al.* which demonstrate its superiority over a number of alternatives and its well-understood behaviour, I adopt the standard local binary pattern (LBP) representation of individual faces [47]. Using LBP I consider two baseline set representations: (i) a set of LBP exemplars, and (ii) a linear LBP subspace, both of which were also evaluated by Wolf *et al.* The former simply stores all face exemplars (that is, the corresponding LBP vectors), while the latter uses principal component analysis to represent the main modes of the observed exemplar variation; previous work (e.g. [48]) suggests that for individual face sets 6-dimensional subspaces produce good results so this is the dimensionality I adopt too.

I examine two baseline set similarity measures, again motivated by the reports of their good performance in the existing literature. The first of these is the *maximum maximorum* ('max-max') cosine similarity between sets of exemplars $\max_{f_1 \in S_1, f_2 \in S_2} f_1^T f_2 / \|f_1\| / \|f_2\|$ which in the experiments of Wolf *et al.* [29] outperformed a number of alternatives including by a large margin the pyramid match kernel of Graumanand and Darrell [49] and the locality-constrained linear coding (LLC) of Wang *et al.* [50]. The second baseline comparison which I adopt for the comparison of sets represented as linear subspaces is the algebraic method based on the maximum correlation between pairs of vectors lying in two subspaces. This method too performed well in the experiments of Wolf *et al.* [29] as well as a number of other authors [**?** 52]. Thus in summary, the two adopted baseline methods are:

- LBP + *maximum maximorum* set similarity, and
- LBP + maximum correlation between subspaces.

These are used to establish reference performance. They are then employed in the context of several different ways of applying the general principle of quasi-transitivity:

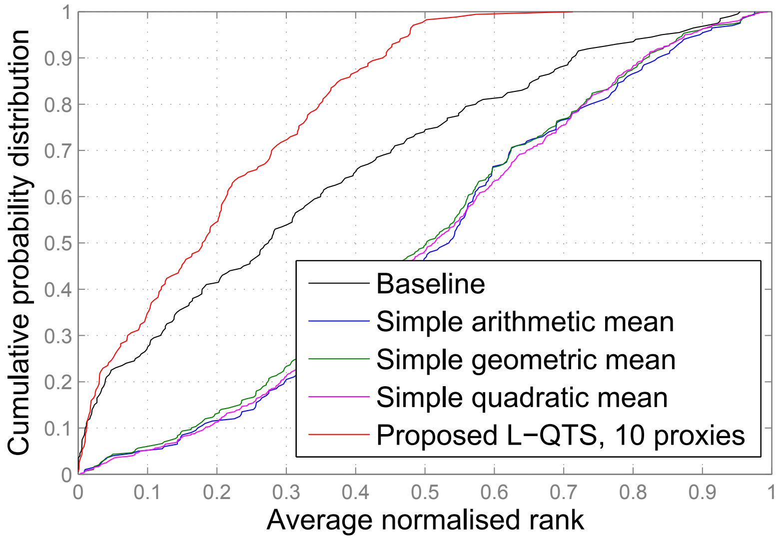
- Simple arithmetic mean based quasi-transitivity,
- Simple geometric mean based quasi-transitivity,
- Simple quadratic mean based quasi-transitivity, and
- Proposed learnt quasi-transitivity (L-QTS)

The first three methods in the list are simple combination rules. In the first of these, the arithmetic mean based quasi-transitivity, two set similarity of dissimilarity measures ρ_{QP} (query-proxy) and ρ_{PT} (proxy-target) are combined by computing their arithmetic mean i.e. $0.5 \times (\rho_{QP} + \rho_{PT})$. Similarly, in the geometric and quadratic mean based methods quasi-transitivity is attempted by computing respectively $\sqrt{\rho_{QP} \times \rho_{PT}}$ and $\sqrt{0.5\rho_{QP}{}^2 + 0.5\rho_{PT}{}^2}$ [53**?**]. The proposed learnt quasi-transitivity (applied atop of both baseline methods) was evaluated using different numbers of proxy sets (1–10) and as detailed in Section 2.4.2, ϵ-SV regression was learnt using the parameter values $\epsilon = 0.4$ and $c = 1000$.

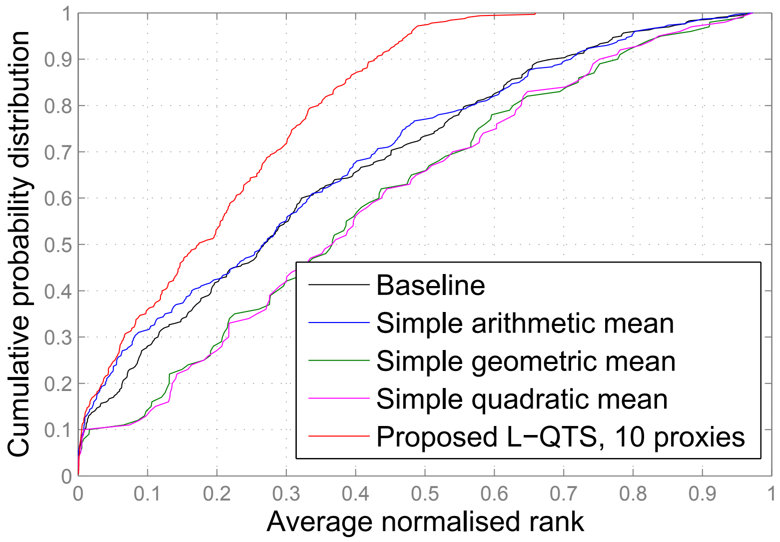
3.4 Evaluation protocol

I train the ϵ-SV regressor using 200 randomly selected sets and their proxies (which are not necessarily in the random 200). In principle there is no reason why the entire database would not be used (recall that no labelling or manual intervention is used whatsoever) but I found that 200 sets were sufficient to gather sufficient training data. Examples are shown in Figure 13; clear patterns are observable both within positive and negative training sets which differ one from another significantly.

The evaluation of the methods described in the previous section was performed by examining all possible retrievals. In other words, I used every set in the database as the query in turn and evaluated the resulting retrieval. To make this feasible I also propose a robust sample selection method so as to reduce the computational demands of the otherwise computationally intensive

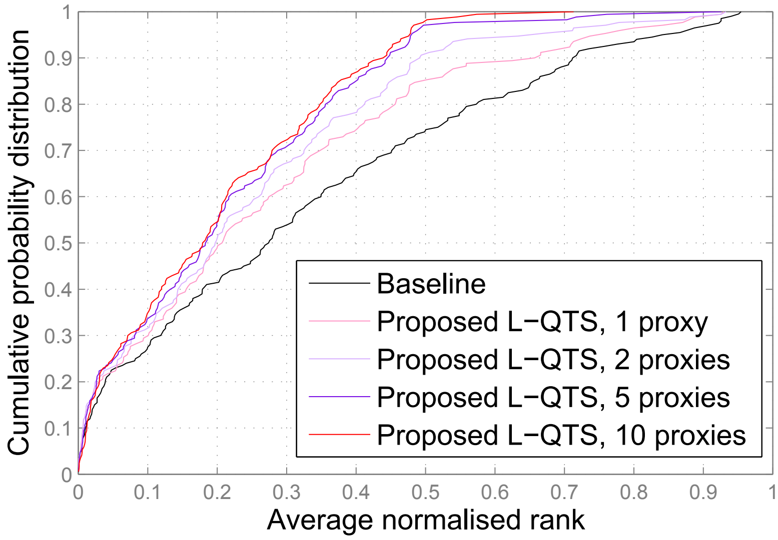


(a) Exemplar baseline, all methods

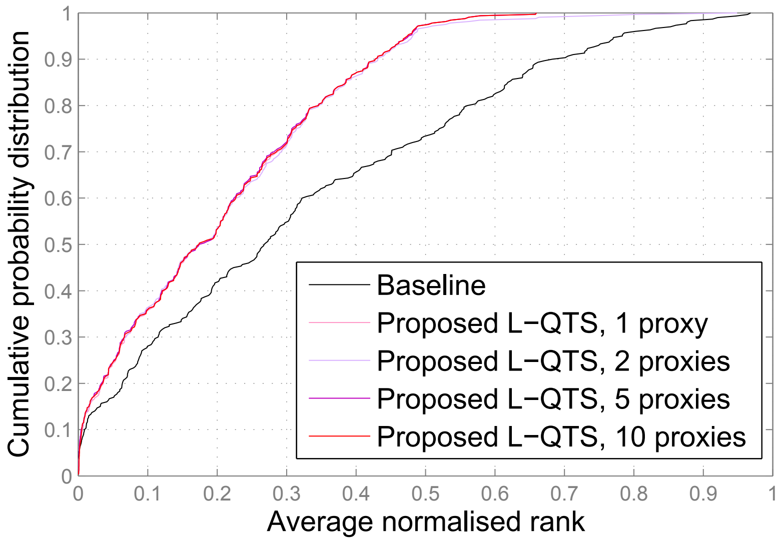


(b) Subspace baseline, all methods

Fig. 11. CDF of the average normalized rank obtained using the exemplar based (a,b) and subspace based (c,d) methods. (a,c) Comparison of the respective baseline approach, the three simple quasi–transitivity estimation methods, and the proposed learnt quasi-transitivity.



(a) Exemplar baseline, proposed



(b) Subspace baseline, proposed

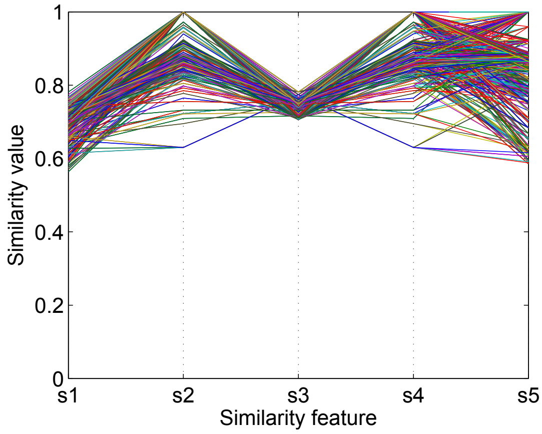
Fig. 12. CDF of the average normalized rank obtained using the exemplar based (a,b) and subspace based (c,d) methods. (b,d) Comparison of the respective baseline approach and the corresponding proposed method for different numbers of proxies.

exemplar based baseline.

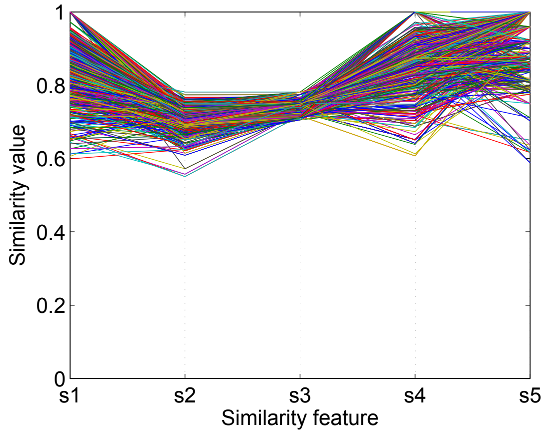
3.4.1 Exemplar baseline: robust sample selection

It is well established by the existing work on face recognition that the appearance of a face is constrained and thus confined to a region of the image space [**?**]. Within this region, which is nonlinear, the appearance variation is mostly approximately smooth – this is sometimes somewhat loosely stated as the face appearance being constrained to a nonlinear appearance manifold [54, 31]. That being said, the range of appearance variation of a person's face within a *single* video typically covers only a portion of the entirety of possible variation. It is a simple yet important observation that even within this range of appearance the underlying manifold is not uniformly sampled, e.g. a person may spend more time in a specific pose than in others. One consequence is that while largely redundant face exemplars of the densely sampled portions of the manifold add little new information about the appearance of the person's face, they can dramatically increase the computational cost of set based comparisons. This is the case for example for face set based comparisons which utilize all sample pairs comparisons such as those based on the *maximum maximorum* similarity (i.e. all pairs maximum similarity) [55] or the *maximum minimorum* distance (a variation of the Hausdorff distance [56]). More worryingly, if a sample voting scheme is used [29], redundant exemplars can unduly affect the result even though they carry little additional information.

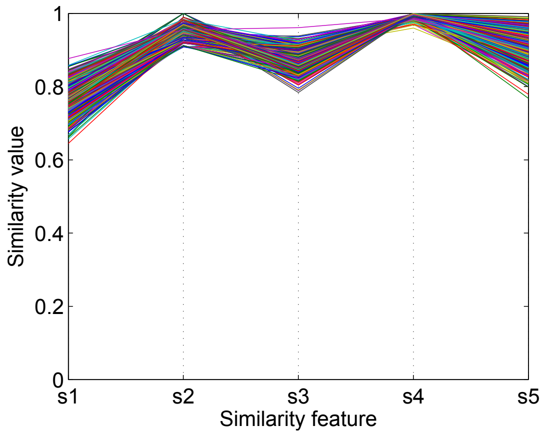
I overcome both of the problems described above by employing a robust sample selection scheme. My starting point is the observation that although the intrinsic dimensionality of the entire face manifold is estimated to be in the range 15–22 [57], the appearance variation exhibited in a typical video clip is typically dominated by a single factor such as face yaw changes; the plot in Figure 8 corroborates this. Led by this insight I employ kernel principal component analysis (KPCA) [58] to project the original face exemplars onto their dominant



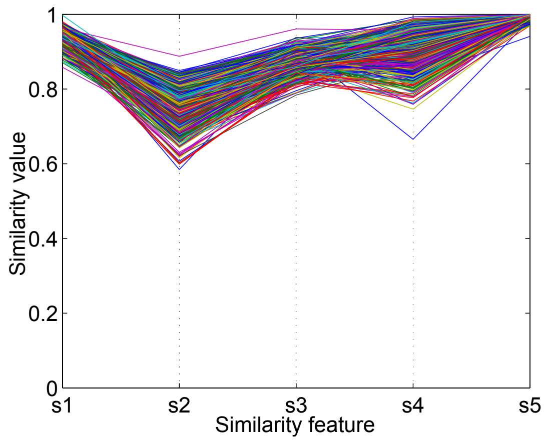
(a) Inter-class transitivity meta-features



(b) Intra-class transitivity meta-features



(c) Inter-class transitivity meta-features



(d) Intra-class transitivity meta-features

Fig. 13. Training data for the exemplar based (a,b) and subspace based (c, d) experiments, in the form of intra-class and inter-class transitivity meta-features. Feature are vectors comprising 5 similarities in (1), and are shown using parallel coordinates [59].

nonlinear principal component, uniformly sample the resulting 1D space between the two projections of the two most extreme exemplars, and finally project them back into the original space. The process is illustrated in Figure 9. The plot in Figure 10 demonstrates that the proposed sample selection does not greatly affect inter-set similarities; a computational improvement of over 2.5 orders of magnitude (approximately 330 times) was achieved.

3.5 Results and discussion

The main set of results from my experiments is summarized in the plots in Figure 11(a) and 11(b) which show the cumulative distribution functions of the ANR achieved for the two baseline methods and different quasi-transitivity approaches. Firstly note that the two baseline methods performed approximately equally well, which is consistent with the previous reports in the literature [29]. The three simple attempts at exploiting quasi-transitivity worsened performance significantly, save for the arithmetic mean based similarity combination for the subspace based baseline which effected neither an improvement nor deterioration. This confirmed the expectation expressed in Section 2.2 that the use of inter-personal similarities only is unlikely to be successful and that a richer set of similarity meta-features is needed instead. This leads us to the proposed method which in both cases effected a major performance improvement over both of the baselines. For example, while the exemplar based baseline produced retrievals with the ANR less than 0.3 in 54.0% of the cases, the corresponding learnt quasi-transitivity did so in 72.5% of the cases (an improvement of 34%). Similarly, while the subspace based baseline produced retrievals with the ANR less than 0.3 in 54.9% of the cases, the corresponding learnt quasi-transitivity did so in 72.8% of the cases. It is particularly interesting to observe in how few cases the proposed method produced bad results (i.e. high ANR) – for both baselines my method achieved ANR lower than 0.5 for over 98% of retrievals. In contrast, the 98% quantile of the baseline methods corresponds to the ANR values of 0.92 and 0.88 for the exemplar and subspace based methods.

The effect of the number of proxies is summarized in Figure 12(a) and Figure 12(b). For both baselines performance improvement is immediately apparent even for the minimum number of a single (i.e. $k_p = 1$) proxy per set. Interestingly, while in the case of the exemplar baseline the performance gradually improves up until $k_p = 5$, staying approximately steady thereafter, the improvement using the subspace based baseline is much more dramatic and reaches its peak (on par with the peak of the exemplar baseline) for $k_p = 1$ already (ANR plots for different k_p are virtually indistinguishable and require significant magnification). Although I are not sure of the exact mechanism that explains this behaviour, it does appear to be linked to the inherent properties of the subspace based baseline which is additionally supported by the observation that the within-class variability of the corresponding training meta-features is significantly smaller than for the exemplar based baseline; see Figure 13.

Let us next turn our attention to the plot in Figure 14(a). It shows the proportion of retrievals (i.e. the empirical estimate of the corresponding probability) which result in at least one correct match being retrieved in the top 100 ranked sets as a function of the total number of target sets in the database which correctly match the query. Plotted as solid blue and red lines are the results obtained using the proposed method (with 10 neighbours used as quasi-transitivity proxies) atop of the exemplar based baseline, and the baseline itself (as expected from Figures 11 and 12, the results for the subspace based method are similar and are thus not included to avoid unnecessary repetition). The plots also show predictions based on the methods' performances for queries in which only a single correct match is present in the entire database. Specifically, starting from the estimate of the probability $p_{1,100}$ of a correct match being retrieved in the top 100 ranked sets using queries where only a single correct match is possible, if different correct matches are ranked independently when k correct matches exist, the probability of at least a single correct match being retrieved in the top 100 is approximately $1 - (1 - p_{1,100})^k$. Since the greatest number of admissible queries (591 individuals in the database have only a single set; clearly these were not meaningful queries for performance evaluation), approximately 48%, has $k = 1$ this is a reasonable estimate to base the prediction on. The estimates are plotted as dashed blue and red lines.

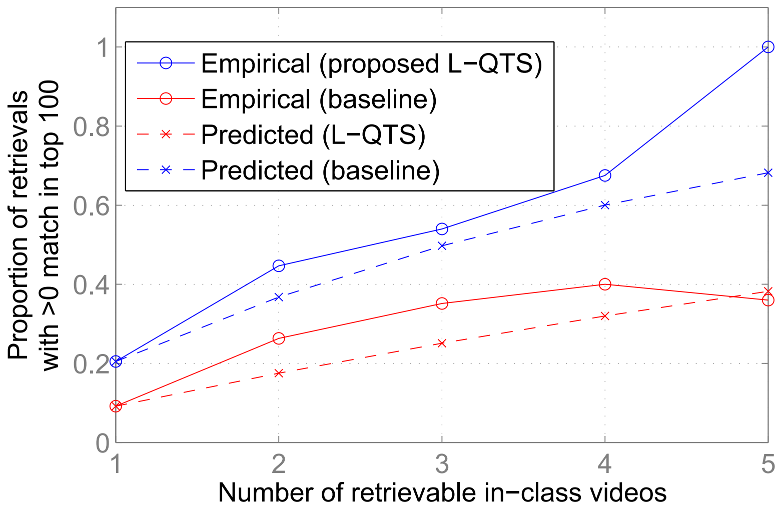
Figure 14(a) reveals interesting insight into the performance of the proposed method. Specifically, note that unlike the empirical plot of the baseline, the empirical plot of the proposed method grows faster with the number of retrievable sets than the corresponding prediction.

This means that the independence assumption underlying the prediction does not hold well, supporting the premise that quasi-transitivity of similarity can be used to improve the retrieval of sets poorly retrieved by the baseline by propagating information from similarly looking individuals or sets of the same person which are acquired in less challenging conditions.

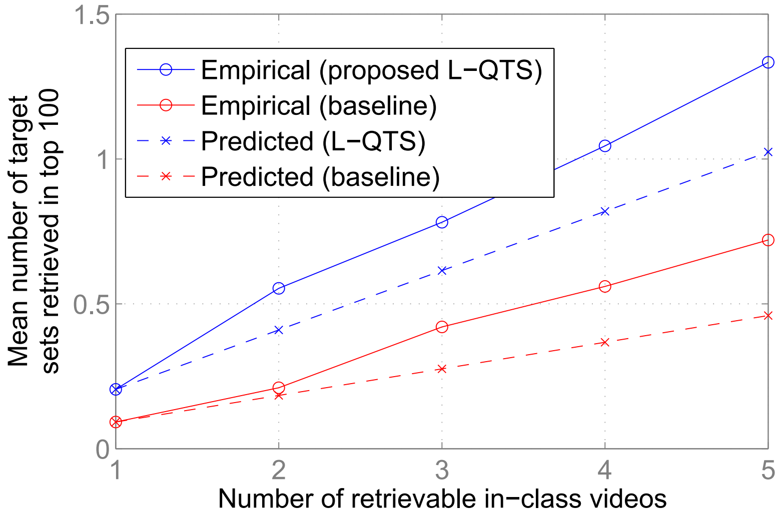
Lastly, Figure 14(b) shows the average number of correct matches retrieved in the top 100 ranked sets as a function of the total number of target sets in the database which correctly match the query. As before the plots also show the corresponding predictions based on the methods' performances for queries in which only a single correct match is present in the entire database. Starting from $n_{1,100}$ the average number of correct matches retrieved in the top 100 ranked sets using queries where only a single correct match is possible, if different correct matches are ranked independently when k correct matches exist, the expected number of correct matches in the top 100 is approximately $k \times n_{1,100}$. The improvement effected by the proposed method is again consistent and significant.

4 Summary and conclusions

In this paper I revisited the challenge widely seen as *the* central problem of face recognition: certain individuals, especially under particular imaging conditions, exhibit a high degree of similarity in appearance. Countering the general consensus across the face recognition community, as well as intuition, I demonstrated that in some contexts – in particular, when the task is that of identity based retrieval from large and highly heterogeneous collections of face image sets – inter-personal similarity can be used to advantage, i.e. it can be utilized to effect an *improvement* in recognition performance. The idea is based on a statistical property of data at a large scale in the form of what I termed quasi-transitivity. I formalized this principle and, to demonstrate its effectiveness, described a specific framework that makes use of it. In particular, I described a meta-algorithm which can be employed with any baseline set matching



(a) Match probability within rank-100



(b) Match number within rank-100

Fig. 14. Rank-100: (a) probability of a correct match being retrieved, and (b) number of correct matches retrieved, vs. number of matches in the database.

algorithm to improve its performance. The baseline method is used to extract meta-features which describe relationships between face sets in the database, which are in turn utilized to learn the form of the corresponding quasi-transitivity function. To facilitate this I also described a general method for automatic extraction of training data from a large, unlabelled corpus. Finally, using a realistic, real-world data set I demonstrated the effectiveness of the introduced ideas empirically. My analysis shows impressive performance, thereby opening a breadth of possible avenues for future research. In particular I would encourage alternative approaches which make use of the concept of quasi-transitivity.

A Extended canonical correlation analysis based baseline

Recall that the key idea behind Extended Canonical Component Analysis (E-CCA) is to bridge the gap between subspace and probability density based representations of within set variability. The aim is to get the best of both worlds, so to speak, by combining seamlessly the advantages of both. More specifically, the main disadvantages of subspace representations lie in the need to make a hard decision on the possible loci of the corresponding patterns, and in turn, the discarding of all second order statistics. On the other hand, probability density based representations suffer from their over-reliance on the statistical representativeness of training data which is an assumption all but universally violated in practical applications of face recognition.

The extraction of meta-features when E-CCA is adopted as a baseline bears a lot of similarity to that of subspace based maximum correlation baseline described previously in Section 2.4.1. As before, meta-feature training data is obtained using only face set pairs which are now represented by the corresponding covariance matrices. To extract training transitivity meta-features which correspond to same identity query-target comparisons, all reference set exemplars f_{qt} iterate through and used to obtain f_{tq} and f_{pq} by anisotropically scaling them

them using respectively the reference and proxy covariances (as in the original work [33]):

$$f_{tq} = \frac{1}{|\Sigma_q|}\mathbf{\Sigma}_q f_{qt} = \frac{1}{|\Sigma_q|}\mathbf{V}_q \Lambda_q \mathbf{V}_q^T f_{qt}, \text{ and} \tag{A.1}$$

$$f_{pq} = \frac{1}{|\Sigma_q|}\mathbf{\Sigma}_p f_{qt} = \frac{1}{|\Sigma_p|}\mathbf{V}_p \Lambda_p \mathbf{V}_p^T f_{qt}. \tag{A.2}$$

The most similar modes of variation, giving f_{tp} and f_{pt} are obtained as per the original work, using eigen-decomposition:

$$f_{tp} = \text{eigv}(\Phi_{pt}, 1), \text{ and} \tag{A.3}$$

$$f_{pt} = \text{eigv}(\Phi_{tp}, 1), \tag{A.4}$$

where $\Phi_{pt} = \sqrt{\Sigma_q}\sqrt{\Sigma_t}$, $\Phi_{tp} = \sqrt{\Sigma_t}\sqrt{\Sigma_p}$, and eigv$(\mathbf{M}, k)$ the k-th eigenvector of $\mathbf{M}$.

The extraction of training transitivity meta-features which correspond to differing identities query-target comparisons proceeds in exactly the same manner, with the difference that it is proxy set exemplars that are iterated through as f_{qt} (as before also taken to be f_{qp}). A single pair of reference and proxy sets gives us n_r 'positive' training transitivity meta-features, where n_r is the number of faces in the reference set, and n_p 'negative' training transitivity meta-features, where n_p is the number of faces in the proxy set. A single pair of reference and proxy sets thus gives us n_p 'negative' training transitivity meta-features.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2004.

[2] Y. Gao, J. Ma, and A. L. Yuille. Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. *IEEE Transactions on Image Processing*, 26(5):2545–2560, 2017.

[3] J. Fan and O. Arandjelović. Employing domain specific discriminative information to address inherent limitations of the LBP descriptor in face recognition. *In Proc. IEEE International Joint Conference on Neural Networks*, 2018.

[4] Z. Dong, C. Jing, M. Pei, and Y. Jia. Deep CNN based binary hash video representations for face retrieval. *Pattern Recognition*, 2018.

[5] T. Kanade. *Picture Processing System by Computer Complex and Recognition of Human Faces.* PhD thesis, Kyoto University, 1973.

[6] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[7] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[8] T. Fromherz, P. Stucki, and M. Bichsel. A survey of face recognition. *MML Technical Report.*, (97.01), 1997.

[9] G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. *In Proc. European Conference on Computer Vision*, 2:581–595, 1998.

[10] S. Edelman and A. J. O'Toole. Viewpoint generalization in face recognition: The role of category-specific processes. *Computational, geometric, and process perspectives on Facial cognition: Contexts and Challenges*, 1999.

[11] The Register. Face recognition useless for crowd surveillance. *The Register, 27 September*, September 2001.

[12] Boston Globe. Face recognition fails in Boston airport. July 2002.

[13] A. S. Tolba, A. H. El-Baz, and A. A. El-Harby. Face recognition: A literature review. *International Journal of Signal Processing*, 2(2):88–103, 2006.

[14] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[15] J. B. Wilmer, L. Germine, C. F. Chabris, G. Chatterjee, M. Williams, E. Loken, K. Nakayama, and B. Duchaine. Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, 107(11):5238–5241, 2010.

[16] O. Arandjelović. Colour invariants under a non-linear photometric camera model and their application to face recognition from video. *Pattern Recognition*, 45(7):2499–2509, 2012.

[17] K.-K. Huang, D.-Q. Dai, C.-X. Ren, Y.-F. Yu, and Z.-R. Lai. Fusing landmark-based features at kernel level for face recognition. *Pattern Recognition*, 63:406–415, 2017.

[18] T. Pei, L. Zhang, B. Wang, F. Li, and Z. Zhang. Decision pyramid classifier for face recognition under complex variations using single sample per person. *Pattern Recognition*, 64:305–313, 2017.

[19] Y.-F. Yu, D.-Q. Dai, C.-X. Ren, and K.-K. Huang. Discriminative multi-scale sparse coding for single-sample face recognition with occlusion. *Pattern Recognition*, 66:302–312, 2017.

[20] X. Yu, Y. Gao, and J. Zhou. Sparse 3D directional vertices vs continuous 3D curves: efficient 3D surface matching and its application for single model face recognition. *Pattern Recognition*, 65:296–306, 2017.

[21] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, and D. Yue. Learning robust and discriminative low-rank representations for face recognition with occlusion. *Pattern Recognition*, 66:129–143, 2017.

[22] X. Wu, Q. Li, L. Xu, K. Chen, and L. Yao. Multi-feature kernel discriminant dictionary

learning for face recognition. *Pattern Recognition*, 66:404–411, 2017.

[23] Z. Fan, D. Zhang, X. Wang, Q. Zhu, and Y. Wang. Virtual dictionary based kernel sparse representation for face recognition. *Pattern Recognition*, 76:1–13, 2018.

[24] I. Schlag and O. Arandjelović. Ancient Roman coin recognition in the wild using deep learning based recognition of artistically depicted face profiles. *In Proc. IEEE International Conference on Computer Vision*, pages 2898–2906, 2017.

[25] A. T. Lopes, E. de Aguiar, A. F. de Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[26] Y. Li, G. Wang, L. Nie, Q. Wang, and W. Tan. Distance metric optimization driven convolutional neural network for age invariant face recognition. *Pattern Recognition*, 75:51–62, 2018.

[27] J. Tang, Z. Li, and X. Zhu. Supervised deep hashing for scalable face image retrieval. *Pattern Recognition*, 75:25–32, 2018.

[28] R. He, T. Tan, L. Davis, and Z. Sun. Learning structured ordinal measures for video based face recognition. *Pattern Recognition*, 75:4–14, 2018.

[29] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011.

[30] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–504, 2011.

[31] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[32] O. Arandjelović. Hallucinating optimal high-dimensional subspaces. *Pattern Recognition*, 47(8):2662–2672, 2014.

[33] O. Arandjelović. Baseline fusion for image an pattern recognition – what not to do

(and how to do better). *The Journal of Imaging (special issue on Computer Vision and Pattern Recognition)*, 3(4), 2017.

[34] O. Arandjelović. Discriminative extended canonical correlation analysis for pattern set matching. *Machine Learning*, 94(3):353–370, 2014.

[35] O. Arandjelović. Learnt quasi-transitive similarity for retrieval from large collections of faces. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4883–4892, 2016.

[36] W. Rieutort-Louis and O. Arandjelović. Description transition tables for object retrieval using unconstrained cluttered video acquired using a consumer level handheld mobile device. *In Proc. IEEE International Joint Conference on Neural Networks*, pages 3030–3037, 2016.

[37] K. Lee, M. Ho, J. Yang, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

[38] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. *In Proc. Asian Conference on Computer Vision*, 2:709–720, 2010.

[39] O. Arandjelović. Recognition from appearance subspaces across image sets of variable scale. *In Proc. British Machine Vision Conference*, 2010. DOI: 10.5244/C.24.79.

[40] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[41] B. Schölkopf and A. Smola. *Learning with kernels.* MIT Press, Cambridge, MA, 2002.

[42] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.

[43] T. Deselaers, D. Keysers, and H. Ney. Classification error rate for quantitative evaluation of content-based image retrieval systems. *In Proc. IAPR International Conference on Pattern Recognition*, 2:505–508, 2004.

[44] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw

Hill, New York, 1983.

[45] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. *In Proc. IEEE International Conference on Computer Vision*, 2:1470–1477, 2003.

[46] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, pages 861–874, 2006.

[47] D. R. Parker, T. D. Ross, and S. C. Gustafson. Receiver operating characteristic and confidence error metrics for assessing the performance of automatic target recognition systems. *Optical Engineering*, 44(9):097202–097202, 2005.

[48] M. Heikkilä, M. Pietikäinen, and C. Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425–436, 2009.

[49] O. Arandjelović. Making the most of the self-quotient image in face recognition. *In Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, 2013. DOI: 10.1109/FG.2013.6553708.

[50] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. *In Proc. IEEE International Conference on Computer Vision*, 2:1458–1465, 2005.

[51] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3360–3367, 2010.

[52] O. Arandjelović and R. Cipolla. Face set classification using maximally probable mutual modes. *In Proc. IAPR International Conference on Pattern Recognition*, pages 511–514, 2006.

[53] O. Arandjelović, R. I. Hammoud, and R. Cipolla. Thermal and reflectance based personal identification methodology in challenging variable illuminations. *Pattern Recognition*, 43(5):1801–1813, 2010.

[54] Y. Haitovsky. A note on the maximization of $\bar{R}^2$. *The American Statistician*, 23(1):20–

21, 1969.

[55] O. Arandjelović. Weighted linear fusion of multimodal data – a reasonable baseline? *In Proc. ACM Conference on Multimedia*, pages 851–857, 2016.

[56] O. Arandjelović. Unfolding a face: from singular to manifold. *In Proc. Asian Conference on Computer Vision*, 3:203–213, 2009.

[57] Y. M. Lui and J. R. Beveridge. Grassmann registration manifolds for face recognition. *In Proc. European Conference on Computer Vision*, 2:44–57, 2008.

[58] T. Cour, B. Sapp, A. Nagle, and B. Taskar. Talking pictures: Temporal grouping and dialog-supervised person recognition. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[59] E. P. Vivek and N. Sudha. Robust hausdorff distance measure for face recognition. *Pattern Recognition*, 40(2):431–442, 2007.

[60] M. B. Lewis. Face-space-R: towards a unified account of face recognition. *Visual Cognition*, 11(1):29–69, 2004.

[61] B. Schölkopf, A. Smola, and K. Müller. *Advances in Kernel Methods – SV Learning*, chapter Kernel principal component analysis., pages 327–352. MIT Press, Cambridge, MA, 1999.

[62] J. Heinrich and D. Weiskopf. State of the art of parallel coordinates. *Eurographics*, pages 95–116, 2013.