Multi-source data inference for object association

By Yogesh Raja, Shaogang Gong and Tao Xiang

Vision Semantics Limited, Queen Mary, University of London
327 Mile End Road, London E1 4NS

Abstract

Effective data fusion from multiple independent sources can be a valuable tool for various defence related scenarios such as multi-camera visual surveillance. Such data can often be huge and widely disparate in nature, with informative associations often being not only previously unknown, but also highly context dependent and sparse. In an example multi-camera visual surveillance scenario, we describe methods for utilising three computational methods for more effectively performing multi-source data association: (1) the use of relative ranking in the form of a method known as RankSVM, replacing the conventional Bhattacharyya distance method for comparing data from two different sources; (2) learning spatio-temporal profiles that model associations between sources in a multi-source informational domain; and (3) the use of interactive data mining to quickly and effectively guide systems to useful and/or previously unknown associations. These methods are applied as part of a Multi-Camera Tracking (MCT) computer vision system for tracking people across disjoint camera views in a multi-camera environment. We describe these components in detail and experimentally demonstrate their effectiveness by comparing the results of the system with exhaustive manually-labelled ground truth for the benchmark i-LIDS dataset. The results show that the combination of these three key components permits the system to operate more quickly and effectively with a significantly higher success rate over previous approaches.

1. Introduction

The "fusion" of information from multiple independent sources, which we refer to as the problem of multi-source data association, can be a valuable tool for various defence related scenarios such as visual surveillance across multi-camera environments or the analysis of intelligence gathered from multiple independent sources of different types of information at different places and times. Such collections of data can often be huge and widely disparate in nature, with informative associations often being highly context dependent and sparse, resulting in the proverbial search for a needle in a haystack. Furthermore, it is not always possible to determine what associations are important prior to searching for them. While various techniques such as the use of prior models or human interaction can help to narrow the search space or guide search with the benefit of human experience, they only offer limited benefits when used in isolation. Within the context of an example multi-camera visual surveillance scenario, we describe a combined Multi-Camera Tracking (MCT) system for utilising three computational methods for more effectively performing multi-source data association: (1) the use of "relative

ranking" rather than absolute metric distance based matching for comparing data from two different sources, offering context-dependent flexibility and softening the pitfalls that result from hard decision based approaches; (2) learning activity profiles which dynamically model the associational structures between sources in a multi-source informational domain; and (3) the use of human interaction as part of a data mining approach to quickly and effectively guide systems to useful connections and associations including previously unexpected ones. This system is deployed as part of a computer vision system for tracking people across disjoint camera views in a multi-camera airport environment. Specifically: (1) we use a method known as RankSVM for ranking rather than classifying similarities between the appearance features of detected individuals in video across different camera views, thereby facilitating greater robustness against incorrect hard decisions; (2) we dynamically learn multi-modal spatio-temporal probability distribution profiles between camera views to reflect the expected movement patterns of people across views and significantly narrow the search space of potential matches by acting as a prior probability; and (3) we permit feedback from users to enable the system to iteratively find more correct matches and in the process discover previously unexpected but potentially high-value associations which may then be incorporated into the search process. We describe these components in detail and experimentally demonstrate their effectiveness by comparing the results of using this system with exhaustive manually-labelled ground truth for the benchmark i-LIDS dataset (Home Office Scientific Development Branch). The results show that combining these three modules permits the system to operate effectively and quickly with a significantly higher success rate over previous approaches.

2. Framework

The design of our multi-camera tracking system begins with computing some feature descriptors. More precisely, the system employs a person detector (Felzenszwalb, Girshick, McAllester & Ramanan (2010)) to automatically generate bounding boxes in video frames corresponding to people. For each bounding box, we extract a feature vector to model appearance. These consist of concatenated histograms of 29 different chromatic and geometric features, including Red-Green-Blue (RGB), Hue+Saturation and YCrCb chromatic features along with Gabor wavelet responses at eight different scales and orientations as well as thirteen differently parameterised Schmid filter responses (Schmid (2001)). Each image patch corresponding to a person is split into six equal horizontal segments, separate histograms are generated for each before concatenation into one. Hence, given 16 bins for each histogram, we have a 2784-dimensional feature vector per image patch which is then normalised and used as a descriptor. We employ a common multi-target tracking technique to locally group detections in different frames as likely belonging to the same person. This process yields *tracklets* encompassing multiple frames. These tracklets, rather than individual detections, are matched across cameras for re-identification.

2.1. *Re-identification by ranking*

A technique is developed for ranking matches between tracklets from different cameras (Prosser, Zheng, Gong & Xiang (2010)). To that end, a model is trained from a data set of pairs of feature vectors from single detections of the same person taken from different cameras. More precisely, given a training set of m samples $X = (\mathbf{x}_i, y_i)_{i=1}^{m}$ where $\mathbf{x}_i \in R^d$ is a feature vector for a specific individual and y_i a corresponding label, and given the feature vectors $\mathbf{x}_j^+$ from the same training set X corresponding to the same person from

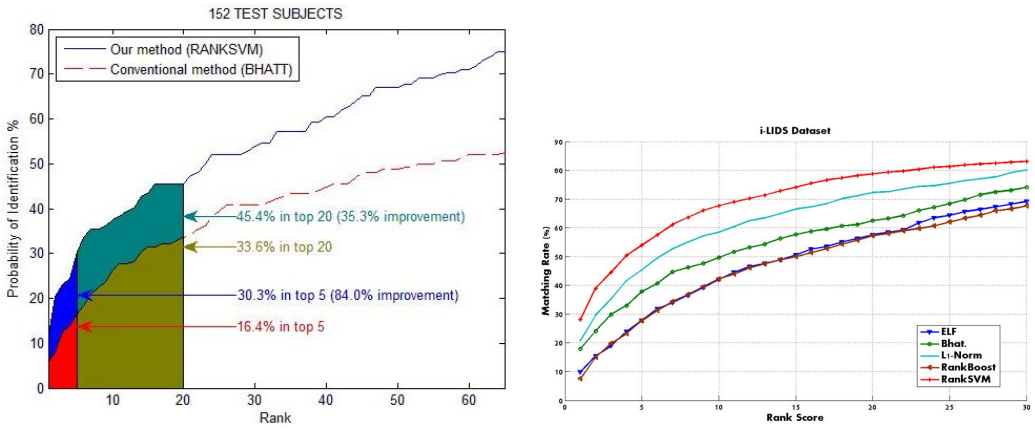
another view (called *relevant* feature vectors) along with $\mathbf{x}_j^-$ corresponding to different people (called *irrelevant* feature vectors), we learn a ranking function δ to rank vector pair similarity such that $\delta(\mathbf{x}_i, \mathbf{x}_j^+) > \delta(\mathbf{x}_i, \mathbf{x}_j^-)$. This takes the form of a support vector machine (SVM) known as RankSVM (Joachims (2002), Chapelle & Keerthi (2010)). The RankSVM model is characterised by a linear function w for ranking matches between two feature vectors as $\delta(\mathbf{x}_i, \mathbf{x}_j) = w^T|\mathbf{x}_i - \mathbf{x}_j|$. Given a feature vector $\mathbf{x}_i$, the required relationship between relevant and irrelevant feature vectors is $w^T(|\mathbf{x}_i - \mathbf{x}_j^+| - |\mathbf{x}_i - \mathbf{x}_j^-|) > 0$, i.e. the ranks for all correct matches are higher than the ranks for incorrect matches. Accordingly, given $\hat{\mathbf{x}}_s^+ = |\mathbf{x}_i - \mathbf{x}_j^+|$ and $\hat{\mathbf{x}}_s^- = |\mathbf{x}_i - \mathbf{x}_j^-|$ and the set $P = \{(\hat{\mathbf{x}}_s^+, \hat{\mathbf{x}}_s^-)\}$ of all pairwise relevant difference vectors required to satisfy the above relationship, a corresponding RankSVM model can be derived by minimising the objective function $\frac{1}{2}||w||^2 + C\sum_{s=1}^{|P|}\xi_s$ with the constraints $w^T(\hat{\mathbf{x}}_s^+ - \hat{\mathbf{x}}_s^-) \geq 1 - \xi_s$ for each $s = 1, ..., |P|$ and restricting all $\xi_s \geq 0$. C is a parameter for trading margin size against training error. Whilst such a single-model ranking approach can be computationally prohibitive with larger and larger training sets due to the resulting rapid increase in the size of P, a solution can be adopted, known as Ensemble-RankSVM (Prosser, Zheng, Gong & Xiang (2010)), for boosting so to combine multiple RankSVM models generated from subsets of the training data with comparable results while avoiding the drastic increase in computational resources.

2.2. *Space-time profiling*

Each camera view is decomposed automatically into regions, across which different spatio-temporal activity patterns are observed (Loy, Xiang & Gong (2009)). Let $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ denote the two regional activity time series observed in the ith and jth regions respectively. Cross Canonical Correlation Analysis (xCCA) is employed to measure the correlation of two regional activities as a function of an unknown time lag τ applied to one of the two regional activity time-series. We denote $\mathbf{y}(t) = \mathbf{x}_j(t + \tau)$ and omit the symbol t for brevity. At each time delay index τ (or each shifting step), xCCA finds two sets of optimal basis vectors $\mathbf{w}_{x_i}$ and $\mathbf{w}_y$ for $\mathbf{x}_i$ and $\mathbf{y}$ such that the correlation of their projections onto the basis vectors are mutually maximised. Let linear combinations of canonical variates be $x_i = \mathbf{w}_{x_i}^\mathsf{T}\mathbf{x}_i$ and $y = \mathbf{w}_y^\mathsf{T}\mathbf{y}$. We wish to estimate the canonical correlation $\boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau)$ (Loy, Xiang & Gong (2009)). The time delay that maximises the canonical correlation between $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ is then computed as: $\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j} = \text{argmax}_\tau \frac{\sum^\Gamma \boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau)}{\Gamma}$, where $\Gamma = \min(\text{rank}(\mathbf{x}_i), \text{rank}(\mathbf{x}_j))$. When matching candidate tracklets between the two regions, only those with corresponding time delay less than $\alpha\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j}$ (α is a constant factor) are considered.

2.3. *User feedback*

Our system employs a "man-in-the-loop" mechanism for providing iterative feedback to the system in order to refine search. Given a new query Q_0, each of the resulting J_0 matches $M_0 = \{m_0(j_0)\}, j_0 = 1..J_0$ can be tagged as correct or incorrect by the user, yielding a set R_1 of K_1 indices for correct flags $R_1 = \{r_{k_1}\}$, $k_1 = 1..K_1$, and a set W_1 of L_1 indices for incorrect flags $W_1 = \{w_{l_1}\}$, $l_1 = 1..L_1$, $R_1 \cap W_1 = \emptyset$, $K_1 + L_1 \leq J_0$. The set $C_1 = \{m_0(r_{k_1})\}$ is then used to initiate a separate query Q_1 to yield J_1 new matches $M_1 = \{m_1(j_1)\}, j_1 = 1..J_1$. These are then combined with the results from the initial query minus the incorrect flagged matches $\hat{M}_1 = \{M_0 \cup M_1\} \setminus \{m_0(w_{l_1})\}$ and the cycle of feedback repeated as many times as required. After n feedback cycles, we have the aggregate pool of matches flagged as correct by the user over all previous feedback iterations $\hat{C}_n = C_1 \cup C_2 \cup ... \cup C_n = \{m_0(r_{k_1})\} \cup \{m_1(c_{k_2})\} \cup ... \cup \{m_{n-1}(c_{k_n})\}$. This set



(a) (b)

FIGURE 1. (a) Cumulative Matching Characteristic (CMC) results as tested on the 152 manually-labelled people. A comparison is also made between the performance using our RankSVM method and the conventional Bhattacharyya distance (BHATT). See text for details. (b) Cumulative Matching Characteristic (CMC) curves comparing our RankSVM approach against all other models currently employed in the literature. This far more controlled experiment was conducted on "cleaner", manually extracted image samples for evaluating matching rates. Again, it is verified that RankSVM outperforms all others.

constitutes the final associated evidence. In our multi-camera tracking application, the matched tracklets may be chronologically ordered and the corresponding video segments combined to produce a reconstruction of the target's movement through the multi-camera environment across different locations.

3. Evaluation

We carried out an extensive systematic evaluation of the performance of our system using the benchmark i-LIDS multi-camera video data (Home Office Scientific Development Branch). The RankSVM model was trained using a set of 208 random image pairs from a training video set *independent* from the test sets selected for evaluation of the system.
Overall Match Characteristics – Figure 1(a) shows the Cumulative Match Characteristic (CMC) of the system as tested on the 152 evaluation targets. A CMC curve shows for any given rank position the likelihood that the highest ranked correct match returned from a query in an arbitrary camera view will lie at that rank position or higher. The solid blue line represents the performance when using our dynamic feature-ranking based approach (RANKSVM). The dashed red line shows the performance of a typical conventional feature vector comparison approach, i.e. Bhattacharyya distance (BHATT) used by other methods. It can be seen that RankSVM results in a 30.3% chance of the best correct match lying in the top 5 ranked results as opposed to 16.4% for Bhattacharyya distance, an improvement of 84%. The probability of the best correct match lying in the top 20 is 45.5% for RankSVM, a 35.3% improvement over the 33.6% shown for Bhattacharyya distance. For clarity, the graph only shows the CMC curves up to Rank 65. However, RankSVM overall demonstrates around a 77% probability of the best ranked correct match lying in the top 100 ranks, as opposed to under 57% for Bhattacharyya distance (an improvement of over 36%).

Figure 1(b) depicts a CMC graph illustrating the result of a more controlled lab experi-

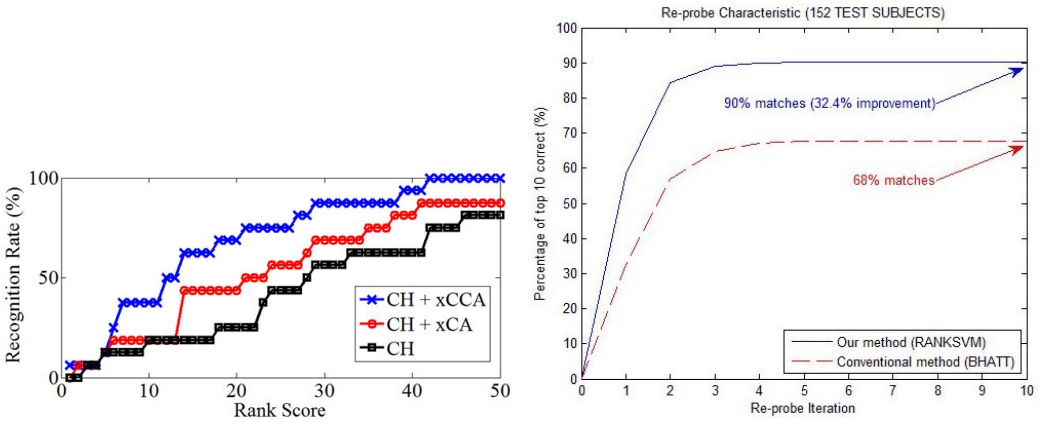


FIGURE 2. (a) The effect of space-time profiling on matching performance. (b) Graph comparing the number of correct matches in the top 10 ranks per re-probe iteration. For RankSVM, it can be seen that we reach around 89% correct matches on average in the top 10 after only 3 re-probe iterations, whereas using Bhattacharyya distance only achieves around 65%. After 10 iterations, we have 90% correct matches in the top ten for our method and 68% for Bhattacharyya distance (an improvement of around 32.4%).

ment conducted on i-LIDS data comparing our RankSVM feature-ranking approach with all other models currently employed in the literature. This experiment is conducted on data sets involving careful manual extraction of image samples of individuals as opposed to the less precise but more real-world relevant, automatically detected and extracted examples used for our MCT system evaluation, with the result that the overall matching rates are higher. Again, it can be verified here that our method (RankSVM) clearly outperforms other approaches.

Effect of Space-Time Profiling – Another key component of this MCT system is the use of space-time contextual knowledge to reduce the search space and increase the likelihood of finding correct matches. These profiles are dynamically learned to reflect the *expected* movements of people throughout a multi-camera environment, thereby providing cues regarding the most likely places and times a nominated target is to appear, either in the future or previously. Figure 2(a) quantifies the benefits of such profiling in more controlled laboratory experiments. It can be seen that the use of colour-histogram based appearance matching alone (labelled CH) without any top-down imposition of space-time expectation to narrow search (Javed, Shafique & Shah (2005)) is outperformed by the addition of a space-time profile (labelled CH + xCCA) (Loy, Xiang & Gong (2009)). The use of space-time profiling improves the results by 175% over the use of matching without profiling. The result also shows that xCCA outperforms the standard Cross Correlation Analysis (xCA) (Kendall & Ord (1990)).

Impact of User Feedback – The user-assisted search paradigm is fundamental to the MCT system. It is designed specifically to assist the user to quickly recover as many correct and relevant match results as possible by iteratively providing feedback to the system about initially recovered matches in order to obtain revised and re-ranked match lists. For each feedback iteration, **accept**ed matches are used by the system to conduct additional searches, called **re-probes**. We evaluate how the percentage of correct matches in the *top ten* ranked results changes with the number of re-probe iterations, from 0 representing

the initial query up to 10 re-probe iterations. Figure 2(b) illustrates this trend. It can be seen that for the system employing our dynamic feature-ranking method (RANKSVM), on average for an arbitrary initial query yielding at least one correct match somewhere in the list, 3 re-probe iterations (rounds of feedback) results in almost 90% of the top ten ranked results in the revised and re-ranked list being correct. Further iterations have little impact on improving the result. On the other hand, for the conventional Bhattacharyya distance method (BHATT), 3 iterations yields on average just over 60%, rising to 68% after 10 iterations. At 10 iterations, this is an improvement of around 32.4%. This graph demonstrates the significant effectiveness of the "man-in-the-loop" user-assisted search paradigm for multi-source data association.

4. Conclusions

We addressed a number of key issues that can have a profound impact on the performance of a multi-source data association system: (a) the use of dynamic feature-ranking; (b) the benefit of space-time profiling as prior knowledge to guide search and improve matching rates; and (c) the effect of the user-assisted search paradigm on the overall effectiveness of the system. Visual matching and re-identification of people in multi-source sensory data is at a fundamental level dynamically contextually-dependent. That is, the most relevant evidence required for discriminating between people vary depending on factors such as the environment and indeed the two entities concerned. Conventional methods using approaches such as Bhattacharyya distance consider all evidence to be universally and equally valid, which circumvents any consideration of context dependence and can significantly blunt data association results. Our approach produces a significant improvement in matching performance when using learned, dynamically context-dependent evidence (feature) ranking, so that different features are selected dynamically when most relevant in different situations. Our current efforts to improve the system are focused on the development of a more extensive learning mechanism to make use of user feedback to: (a) improve initial match rankings by incrementally improving the RankSVM model as part of an active learning mechanism; and (b) update spatio-temporal profiles to reflect changes in movement patterns over time.

REFERENCES

Chapelle, O. & Keerthi, S. 2010 Efficient algorithms for ranking with SVMs. *Information Retrieval* **32**.

Felzenszwalb, P., Girshick, R., McAllester, D. & Ramanan, D. 2010 Object Detection with Discriminatively Trained Part Based Models *PAMI* **32**.

Fraser, A. M. & Swinney, H. L. 1986 Independent coordinates for strange attractors from mutual information. *Physical Review*, **33(2)**, 1134–1140.

Home Office Scientific Development Branch Imagery Library for Intelligent Detection Systems *http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/*.

Javed, O., Shafique, K. & Shah, M. 2005 Appearance modeling for tracking in multiple non-overlapping cameras. *ICCV*, **2**, 26–33.

Joachims, T. 2010 Optimizing search engines using clickthrough data. *Knowledge Discovery and Data Mining*, 133–142.

Kendall, M. & Ord, J. K. 1990 Time Series. *Edward Arnold*.

Loy, C. C., Xiang, T. & Gong, S. 2009 Multi-Camera Activity Correlation Analysis. *CVPR*.

Prosser, B., Gong, S. & Xiang, T. 2008 Multi-camera matching under illumination change over time. *ECCV Workshop on Multi-camera and Multi-model Sensor Fusion*.

Prosser, B., Zheng, W., Gong, S. & Xiang, T. 2010 Person Re-Identification by Support Vector Ranking. *BMVC*.

Schmid, C. 2001 Constructing models for content-based image retrieval. *CVPR*, 39–45.