Human Identity Recognition in Aerial Images

Omar Oreifej
Computer Vision Lab
University of Central Florida
Orlando, FL
oreifej@eecs.ucf.edu

Ramin Mehran
Computer Vision Lab
University of Central Florida
Orlando, FL
ramin@cs.ucf.edu

Mubarak Shah
Computer Vision Lab
University of Central Florida
Orlando, FL
shah@eecs.ucf.edu

Abstract

Human identity recognition is an important yet under-addressed problem. Previous methods were strictly limited to high quality photographs, where the principal techniques heavily rely on body details such as face detection. In this paper, we propose an algorithm to address the novel problem of human identity recognition over a set of unordered low quality aerial images. Assuming a user was able to manually locate a target in some images of the set, we find the target in each other query image by implementing a weighted voter-candidate formulation. In the framework, every manually located target is a voter, and the set of humans in a query image are candidates. In order to locate the target, we detect and align blobs of voters and candidates. Consequently, we use PageRank to extract distinguishing regions, and then match multiple regions of a voter to multiple regions of a candidate using Earth Mover Distance (EMD). This generates a robust similarity measure between every voter-candidate pair. Finally, we identify the candidate with the highest weighted vote as the target. We tested our technique over several aerial image sets that we collected, along with publicly available sets, and have obtained promising results.

1. Introduction

Identity recognition from aerial platforms is a daunting task. The objects of interest, humans for the purpose of this research, have articulated bodies which account for highly variant features in different poses and vanishing details under low quality images. Therefore, previous techniques that heavily relied on image details face significant difficulties.

On the other hand, object identity recognition and object tracking can be closely related problems, since solving tracking implicitly solves identification and solving identification over consecutive frames is actually tracking. However, in tracking, objects are usually considered to have small displacements between observations. Therefore, most tracking techniques such as Mean Shift [4] and

Kalman filter-based tracking make use of this information and search for the tracked objects within small spatial variation limits. Such techniques have proved their efficiency in continuous scenes where disappearances and clutters are minor. However, in the cases of long occlusions, tracking performance considerably decays and it even becomes totally inefficient when discontinuities are inherent in the video. As the problem shifts from solving correspondence in a smooth continuous video to static images with long temporal gaps, all assumptions of the continuous motion models become weak, and the solution resolve to static image based recognition rather than tracking.

To the best of our knowledge, recognition of humans across aerial images has not been directly tackled before. The problem faces the challenges of low quality images, minor availability of details, high pose variations, and the possibility of high density crowds in the same image. Therefore, we provide a robust solution that makes use of specific computer vision tools which work efficiently even under such deteriorated scenarios. For instance, face detection was previously employed for human recognition in high quality photographs; however, it is far from working under aerial views; therefore we employ human detection which has proven efficiency even in low quality aerial imagery. The proposed method overcomes the discontinuity and the information loss in such environment by employing a robust region based appearance matching.

The rest of the paper is organized as follows: In the next section, we present an overview of the related works. Section three provides the problem definition. In section four, we describe the proposed method that we refer to as Weighted Region Matching (WRM). Section five illustrates the results and discusses the experiments. Finally, section six presents the conclusions.

2. Related Work

Not much work has been reported in the literature of object identity recognition, in which a specific individual from

a certain object class is matched using only static image to image comparison. This is mainly because distinct class members have similar shape and appearance, which makes the problem both complex and limited. In [9], Guo *et al.* presented a complete framework for vehicle matching in aerial views. However, their main focus was blob extraction and alignment rather than recognition. In [2], people were recognized in photo albums by employing a Markov Random Field that incorporates both face and clothes similarity potentials. On the other hand, logistic regression was used in [18] to combine similarity scores from the faces and the clothes in order to cluster humans according to their identities. In [8], Gray and Tao proposed using Adaboost to learn the best features for human identification in addition to learn the weak classifier model in the same framework. In [17], people's identities were corresponded across repeated shots of the same scene via pictorial structure that starts also from face detection. Moreover, two human identification methods were proposed in [7]; the first applies a graph based spatiotemporal segmentation to group human pixels that belong to the same fabric. The second method uses decomposable triangulated graphs to segment and correspond the different human body parts. All previous methods are highly dependent on image details to extract features such as faces or body parts, and therefore can only be applied to high quality ground images.

Even though human recognition has not been the focus of the literature, substantial advances in human detection have been reported. In [6], Dalal and Triggs trained a SVM classifier using features of Histograms of Oriented Gradients (HOG) for human detection and localization. Further improvements on HOG were later proposed in [20]. On the other hand, in [19], covariance features were utilized as pedestrian descriptors for a learning algorithm which used Riemannian manifolds to classify pedestrians. Meanwhile, body shape models have also been used for human detection as in [10, 11].

In this work, we identify a target in an aerial image by employing the best tools and features to preprocess the human figures in order to generate two sets of blobs which we weight using graph theory and then correspond through a region-based appearance matching.

3. Problem Definition

We lay the problem as the following: A user is able to identify a target person in some instances from a set of unordered aerial images that were taken for a scene over a short period of time, in a way such that humans maintained their clothing and general appearance. Using only this information, it is required to locate the target in each query image from the set. In an abstract view, we define the problem as a voter-candidate race, where the set of manually
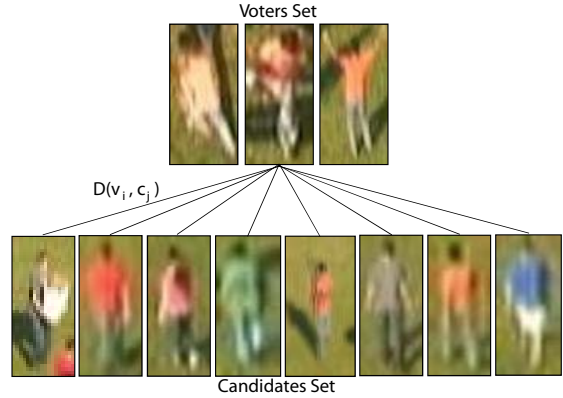

Figure 1. Voters-Candidates. First row shows a possible set of voters for a target, second row shows a possible set of candidates from a certain query image to which the voters need to be matched.

labelled target figures are referred to as voters, and the set of all the human figures in a query image are referred to as candidates. In other words, the problem is to find the candidate that best matches to the voters. Figure 1 shows an example for the voters and candidates.

The underlying challenge of the problem arises from two factors: The low quality of the images, and the pose variation across the set which is due to the changes in camera location and the articulation of the human body. Thus, we show that our proposed method accommodates for the low quality images and is capable of matching in different poses.

4. Weighted Region Matching (WRM)

The proposed matching operation expects as an input a few images where the target has been recognized. Subsequently, the input images and each query image undergo the following preprocessing steps:

- Human Detection.
- Blob Extraction.
- Alignment.

The above steps are explained in the following subsections. The result of these steps is two sets of blobs, the voters and the candidates. If the set of the voters is defined as $V = \{v_i; i = 1...n\}$, and the set of the candidates in a certain query image is defined as $C = \{c_j; j = 1....m\}$, then the probability of blob c_j corresponding to the target is denoted by:

$$P_T(c_j) = \sum_{i=1}^{n} P(c_j|v_i)P(v_i), \qquad (1)$$

where $P(v_i)$ is the voter's prior. If w_i is a weight assigned to voter i, and $D \in [0, 1]$ is the normalized distance between c_j and v_i, then equation 1 can be rewritten in a form similar to a mixture of Gaussians:

$$P_T(c_j) \propto \sum_{i=1}^{n} (exp(-D(c_j, v_i)/\tau)) \times w_i, \qquad (2)$$

where τ is a constant parameter. Hence, in order to solve the recognition problem efficiently, we need to provide a robust representation of the distance between every voter-candidate pair. Moreover, we need to specify the weight of every voter according to its importance in representing the target's specific information.

For further refinement and scene learning, matched candidates with high confidence can be augmented with the voters set V in order to capture more information about the target. The various steps associated with the proposed Weighted Region Matching are illustrated in Figure 2.

4.1. Human Detection and Blob Extraction

Using only static image information, we seek detection and extraction of human blobs. For detection, we train a support vector machine (SVM) classifier based on the HOG descriptor [6] using a dataset of pedestrian images in aerial view. The HOG descriptor captures the most important cues of the human body, such as head and shoulders in good detail. Interestingly, in the case of aerial images, the most observable parts of the body are the head and shoulders. Therefore, Dalal and Triggs' algorithm [6] is a good choice for aerial scenes.

We used 6000 positive images of humans at different scales and poses from a subset of manually labelled images of our aerial dataset, along with 6000 negative examples of the background and non-human objects. A SVM classifier was trained over a subset of 9000 with equal numbers of positive and negative examples. We performed validation using the rest of the dataset. Figure 3 illustrates an example of human detection results.

Human detection outcomes are bounding boxes of pedestrians. However, the background regions contained in the boxes do not provide any information about a specific person. In fact, when the same pedestrian is sighted in different surroundings, the background context causes ambiguity which eventually results in false matchings. Several segmentation methods could be used to separate the regions of the background from the regions of the human's blob; however, we found in our experiments that it is better to estimate one distribution for each of the background and the foreground using a kernel density estimator [12, 15]. Assuming that the human's figure will be centered in the bounding box, we use the center points as initial samples to estimate the foreground PDF, and the border points as initial samples to estimate another PDF for the background. Consequently, we compute pixel probabilities for the background and the foreground and assign every pixel in the bounding box to its most probable distribution. Iteratively,
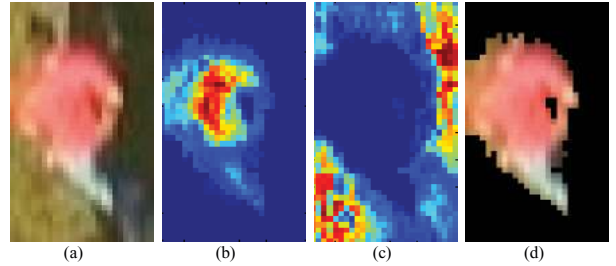

Figure 4. Blob extraction (a) Bounding box obtained from Human Detection. (b) Final probabilities of pixels for the foreground joint kernel density estimation (KDE). Jet color map is used where blue corresponds to the lowest probability and red to the highest. (c) Final probabilities of pixels for the background joint kernel density estimation using Jet color map as well. (d) Extracted blob.

we refine the kernel density estimation based on the newly assigned pixels and then assign pixels based on the refined distribution. We continue this iterative algorithm until foreground-background segmentation stabilizes. An illustrative example is shown in Figure 4.

4.2. Alignment

When comparing blobs, aligning one to the other before matching brings both of them to a unified view and accordingly eliminates the variations from camera orientation and human pose. However, since the size of pedestrians in aerial images is generally very small such that the details of the body are not distinguishable, alignment techniques based on body part detection such as [10] or 3D pose tracking such as [1] do not perform desirably. On the other hand, in the kind of images we are dealing with, edge detection is noisy, and as a result, alignment methods that completely rely on edge detection as in [9, 10] become weak as well.

For the purpose of matching, we are rather interested in a coarse alignment that can capture the general basic pose of the pedestrian under such severe conditions. Thus, we rely on a less detailed model that captures the most visible parts of the body. We use an eight point head, shoulders and torso (HST) model shown in Figure 5.a. The model captures the basic orientation of the upper part of the body. We neglect the body limbs since they are far more articulated, and hence could result in extreme variations; in addition that they tend to vanish in low resolution images.

In order to find the best fit of the HST model over human blobs, we train an Active Appearance Model (AAM) [5] over scaled human images obtained from our dataset. The initial position of the model is vital for AAM to find the correct fit; therefore, we take advantage of the blob region information by using the major and minor axis lengths to initialize the model's size, and the major axis angle to initialize the model's rotation. Figure 5 shows the HST model after fitting over human blobs.

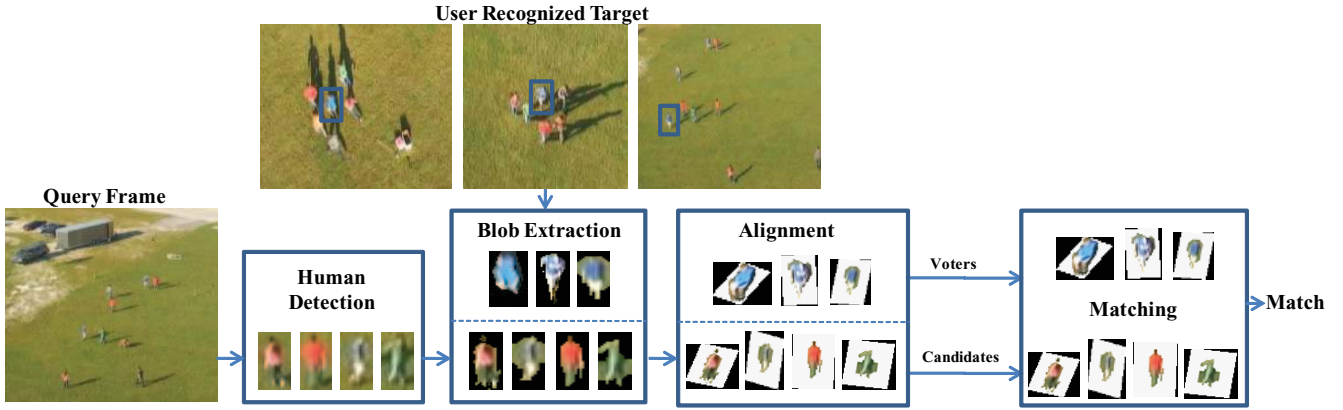By fitting the model over the blobs, we obtain matching

Figure 2. The Weighted Region Matching (WRM) steps. A sample output of detection, extraction, and alignment is presented inside every step box.
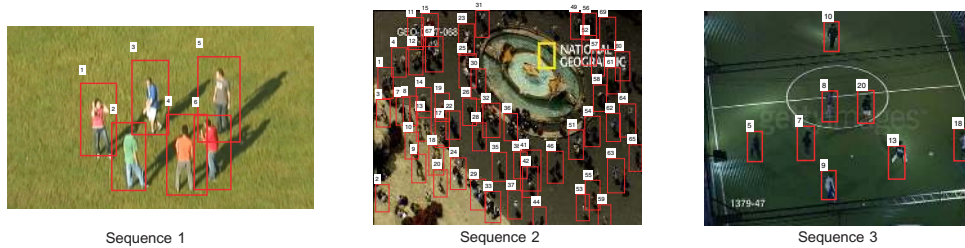

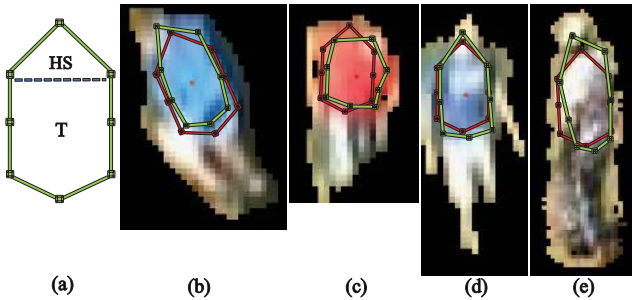Figure 3. Examples of human detection results in three image sets.

Figure 5. The Head-Shoulders-Torso (HST) model fitting. (a) The model points. The upper part (HS) captures the head and shoulders. The lower part (T) captures the torso. (b,c,d,e) Example results of using AAM to fit the HST model over human blobs. The red color shows the model after initialization. The green color shows the final model location after running AAM.

points that we employ to compute a full affine transformation to a desired pose. In our experiments, we align all the blobs to the mean pose generated by the AAM training set. Example alignment results are shown in Figure 6. It is important to notice that the results do not show perfect alignment but are rather quite adequate for matching purposes.

4.3. Measuring the Distance Between Blobs

Given the extracted human blobs, we introduce a measure to effectively distinguish a specific person from others. In order to account for the severe imaging conditions and the articulation of the human body, we propose a matching method that considers a detailed representation of the blob

Figure 6. Example results of alignment. The first row shows the original blobs, the second shows the blobs after alignment. Blobs are closer to common orientation due to alignment.

by treating it as a group of small regions of features. We use Mean Shift with small spatial bandwidth to over-segment each blob into several regions using color features.

In the following, we compute a set of features for each region on a human blob. These features comprises (1) histograms of color values (HSV channels) of the contained pixels, (2) the HOG descriptor of the center of the region. We apply PCA on the feature space and extract the eigen vectors corresponding to the top 30 eigen values, and accordingly project our feature space onto the principle components. The resulting is a set of histograms which represent the regions. It is worth mentioning that we experimented with other types of features but they turned out to be useless; for example, texture did not improve the performance, which is expected since it is blurry in such low quality images.
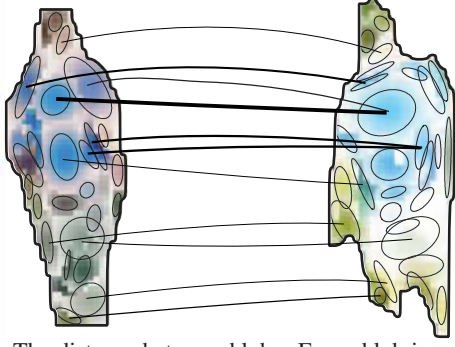
Figure 7. The distance between blobs. Every blob is a set of signatures (regions) where a ground distance relates every two region pair. EMD performs many to many matchings between the blobs' signatures. Connections in the figure represent the EMD flow matrix associations. Some connections are suppressed for clarity.

Using Earth Mover Distance [16, 14] (EMD), the distance between two blobs is computed as the minimum cost of matching multiple regions from the first blob to multiple regions from the second. Having each region represented as a distribution in the feature space (a feature vector), the blob is then represented as a collection of distributions; therefore the problem of matching two blobs becomes the problem of measuring the effort of converting all distributions in one blob to the other.

Using EMD terminology, we refer to every blob as a set, and every region is denoted by a signature and a weight. The ground distance relates every pair of regions. EMD allows many to many associations; which becomes very useful in our problem, since a region from one blob can be related to more than one from the other, depending on the segmentation result. For ground distance, we use Jeffrey-divergence (JD) [14] between every pair of regions. For two distributions $P = \{p_i\}$ and $Q = \{q_i\}$, JD is defined as:

$$JD(P,Q) = \sum_i [p_i \log_2 \left(\frac{p_i}{(q_i + p_i)/2} \right) \\ + q_i \log_2 \left(\frac{q_i}{(q_i + p_i)/2} \right)]. \tag{3}$$

JD is a symmetric and more numerically stable version of Kullback-Leibler divergence (KL). For every region region pair, we also augment their JD distance with the Euclidian distance between their locations within the bounding box. The final ground distance between two regions P and Q is $D_{P,Q} = JD(P,Q) + \alpha \times ED(X_P, X_Q)$, where α is a weighting parameter, ED is the Euclidean distance, and X is the location of the region's centroid within the box. It is worth mentioning that several distance measures were considered, and experiments showed the effectiveness of the proposed combination of distances. Figure 7 illustrates the matching of two blobs using EMD.

4.4. Determining the Voter's Weight

Each region in the voter's blob has certain information about the target; some regions could be noise introduced from the blob extraction or even from the scene itself. Therefore, we rank the collection of input images according to the value of information they carry about the target. In other words, given the set of regions from all voters, $R = \{r_k\}$, we assign a weight for every region such that the most consistent regions are given higher weights, because they are more probable to lead to the target's identity. Consequently, we aggregate the regions' weights for each voter to determine its weight. The PageRank algorithm [3] provides a neat solution for the problem.

PageRank is used to grade websites based on a random walk algorithm which not only gives higher scores to the websites that have more incoming links but also the pages that are referred to from prominent webpages. Therefore, in a graph of connected webpages, the most informative pages are associated with higher ranks. We use the same idea to seek the most informative regions from the set of all voters' regions R. The intuition behind the PageRank weighting is that a voter's region r_i that holds vital information about the target will be consistent across the voters, and accordingly, other voters will contain regions that are close to r_i in the feature space. On the other hand, noisy regions are inconsistent; therefore, they are more likely to be connected to highly separated regions. To perform PageRank, we create an undirected graph $G = (R, E)$, where the regions' set R represents the nodes of the graph, and E is the set of edges connecting each pair of regions. In G, we connect every region from voter i to the K nearest neighbor regions of voter j where $i \neq j$. In addition, we assign a distance to every edge $e_{i,j}$ connecting two regions r_i and r_j which is computed as described in Section 4.3. As a result, PageRank will assign higher weights to prominent regions and degrade noisy regions by assigning them lower weights. Figure 8 demonstrates the weighting process for an example set of 5 voters. It is quite hard to visualize the effect of the weighting on such low quality blobs; therefore, we demonstrate it's effect on synthesized noisy regions overlaid on a set of blobs in Figure 9.

The region rank obtained from PageRank is not directly dependent on the region size. However, the region size rather represents the amount of information in the region; therefore, we define the final weight for a region r_k as: $w_k = w_k^{pr} \times w_k^s$, where w_k^{pr} is the region's normalized PageRank weight and w_k^s is the normalized size of the region. Consequently, the voter's weight is computed as the normalized sum of weights of its regions. Therefore, for a certain voter v_i with s regions, the w_i in Equation 2 is defined as:

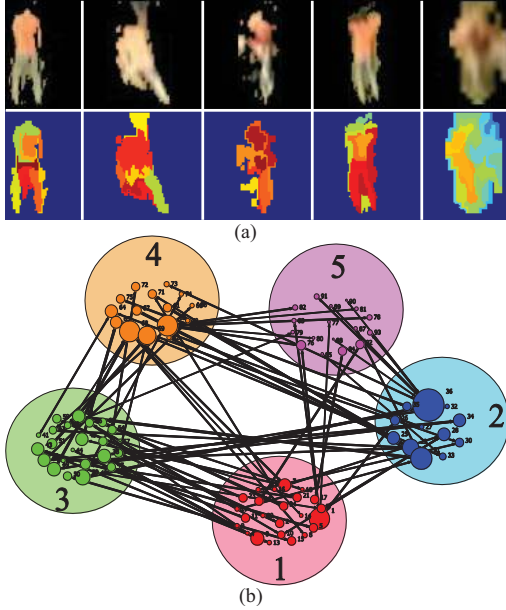$$w_i = \frac{\sum_{k=1}^{s} w_k}{\sum_{i=1}^{n} w_i}. \tag{4}$$

Figure 8. (a) The first row shows a set of five voters for a certain target. The second row shows the regions' weights assigned by PageRank represented in Jet color map. The noisy regions (inconsistent) were assigned lower weights. (b) PageRank Graph for the voters. Outer circles represent voters. Inner circles represent regions, where the circle size corresponds to the region weight. A region of a given voter is connected to k regions of other voters. Some connections are suppressed for clarity.
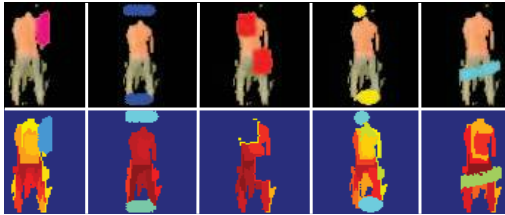

Figure 9. The PageRank Weighting. The first row shows a set of five voters with overlaid synthesized noisy regions. Second row shows the weights assigned by PageRank represented in Jet color map. The noisy regions were assigned the lowest weights.

4.5. Matching

The previous steps generate two sets of extracted and aligned voters and candidates blobs, along with associating every voter-candidate pair with a distance that represents their localized separation in the feature space. In addition, the weighting step ranks every voter based on the significance of its regions in representing the target. Substituting the distances and the weights in equation 2, we compute a probability for every candidate to belong to the target.

The matching result is the computed probabilities. In the simplest setup, the best match should be the candidate with the highest probability. Moreover, a confidence about the actual existence of the target in a query image is inferred from the probability values as well. A low confidence pro-

vides enough evidence for the disappearance of the target in the image.

5. Experiments and Results

We thoroughly tested our proposed method over several challenging image sets mainly taken from a collection of aerial images that we captured using a UAV and a set of human actors. In addition, we tested our method on several images from the web. The dataset is available on our website.

We obtained manual annotations for all the sequences. The overall experiment dataset contained about $6,000$ human bounding boxes. The whole dataset contributed in training the human detection and the alignment processes, in addition to providing a ground truth for WRM testing. For the purpose of isolating the error sources from recognition and detection, all matching experiments were conducted using the true positives of human detection process.

In our experiments, we used the final candidates' probabilities P_T to obtain the following quantitative performance evaluation metrics:

- Precision: The ratio between the number of correct matches and the number of testing query images.
- Mean Average Precision: Since we have only one target in each query image, this measure is defined as the mean percent ranking of the ground truth candidate across all testing query images. The rank is computed for every query image such that the highest rank is assigned to the candidate with the highest P_T. For instance, in a certain query image, if the ground truth candidate had the highest P_T, its Average Precision is $(1/1) \times 100\% = 100\%$. In another query image, if the ground truth candidate was ranked third, its Average Precision is $(1/3) \times 100\% = 33.3\%$. The Mean Average Precision for the two images is $(100 + 33.3)/2 = 66.6\%$.

In order to study the effect of the different components of our recognition system, we conducted our experiments using all possible combinations of components. Figure 10 and 11 summarizes the average results, which were obtained by running the algorithm using all the testing queries for different numbers of voters. It is clear from the results that every component of the system contributes in the overall performance. It's also worth noticing that PageRank weighting actually causes a drop in the performance when applied without Blob Extraction. The reason for this is that the weighting process selects the most consistent regions among the voters, and in this case background will still exist in the boxes and will also be consistent; therefore background regions will be falsely given high weights. The figures also show that both Precision and Mean Average Precision generally improve with the increased number of voters up until a certain limit. This is because adding more
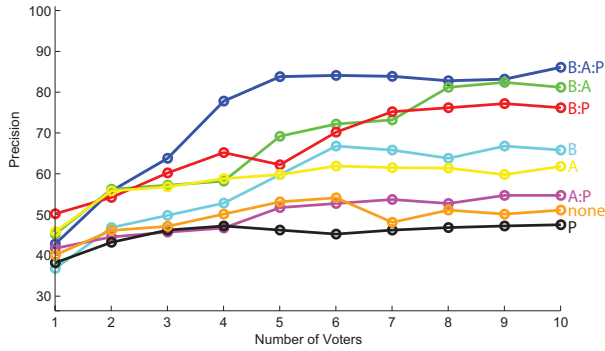
Figure 10. Precision vs Number of voters. Each curve represents a different combination of the system modules coded by the letters on the right, where B stands for Blob Extraction, A for alignment, and P for PageRank Weighting. For example, B:A is the combination of using both Blob Extraction and Alignment but without PageRank Weighting.
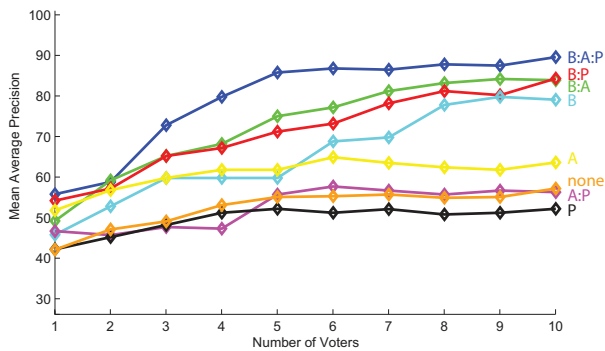

Figure 11. Mean average precision vs Number of voters. The letter code is the same as in figure 10.

prior information about the target increases the probability of capturing it in different poses. Table 1 illustrates example results obtained by running WRM over several image sets.

Since the problem we are dealing with is considered quite novel, there are no directly similar methods to compare our results with. For instance, techniques in [2, 18, 17] are based on face detection, however, faces are not visible from aerial views, and therefore such methods do not fit in this environment.

In view of the fact that solving recognition (matching) over consecutive frames can be used in tracking, we applied the proposed recognition system on a video and compared the results to the state of the art tracking proposed in [13], which applies a cross correlation matching supported by a social behavior-based model of the human's motion. WRM delivered a very close performance as shown in figure 12. The tracking is generally more persistent in high frame rate videos since it obeys the assumption that the object position in the next frame should be close to its position in the current frame, while the recognition does not enforce any motion assumptions. However, recognition becomes more effective in the case of long occlusions where the motion assumptions do not hold. In addition, tracking suffers from

Figure 12. Tracking via recognition for two example humans. The two targets in the black boxes were reidentified by WRM at locations shown in red and yellow dots respectively. Voters were selected at three random frames. Social behavior-based tracking assigns different labels to the same object when it appears again after occlusion, this is shown by the the solid lines with different colors (blue, magenta,cyan) which all correspond to the track of the human on the top left. WRM recognition is free of motion assumptions and therefore maintains consistent labelling and can efficiently match over any set of even separated and unordered images; therefore, it can be used as a robust correspondence measure to join the cluttered tracks.

occasional assignment of new labels to humans who came back to the field of view after being occluded, while WRM deals with such ambiguities quite efficiently. we further decreased the video frame rate in steps and observed the tracking becoming completely useless when frames are highly separated, while our recognition system still worked well. It is important to notice that the proposed system is not a tracking method and should not be compared to tracking systems; however, it could be used as a robust correspondence measure within a tracking framework.

It is rather essential to mention that in the highly crowded scenarios where inter-person occlusions occur frequently, performance of WRM declines. In this case, the detected human bounding boxes contain several merged blobs, and accordingly blob extraction and alignment steps become weak. Nevertheless, WRM is able to find informative regions of the target and assign them high weights by applying the PageRank. Therefore, it still matches reasonably.

6. Conclusions

We have presented a framework for detecting, segmenting, aligning, and recognition of humans viewed from aerial cameras with low resolution and tough conditions. The identity of a target was recovered by detecting and corresponding salient regions of the target's blob. Our recognition system is domain independent since it does not force any motion model or pose restrictions. Experiments showed high accuracy and robustness with mean average precision

Voters	Query Image	Candidates	Matching Result

Table 1. Example Experimental Results. Candidates' column shows a sample from the candidates in the query image. Every candidate is associated with a probability P_T shown on its upper right. The matched candidate is circled.

of 89.6%. In future work, we will be investigating the use of additional heterogenous features.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. *CVPR*, 2004.

[2] D. Anguelov, K. chih Lee, S. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. *CVPR*, 2007.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, 1998.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *CVPR*, 2000.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, 2001.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.

[7] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. *CVPR*, 2006.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with ensemble of localized features. *ECCV*, 2008.

[9] Y. Guo, H. Sawhney, R. Kumar, and S. Hsu. Robust object matching for persistent tracking with heterogeneous features. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[10] M. W. Lee and R. Nevatia. Body part detection for human pose estimation and tracking. *WMVC*, 2007.

[11] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *CVPR*, 2005.

[12] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 1962.

[13] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. Youll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking. *ICCV*, 2009.

[14] J. Puzicha, J. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. *ICCV*, 1999.

[15] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 1956.

[16] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 2000.

[17] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. *BMVC*, 2006.

[18] Y. Song and T. Leung. Context-aided human recognition - clustering. *ECCV*, 2006.

[19] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *PAMI*, 2008.

[20] X. Wang, T. X. Han, and S. Yan. An HOG-LBP Human Detector with Partial Occlusion Handling. *ICCV*, 2009.