Monitoring Interactions

Felix Meißner[1] and Remco C. Veltkamp[1]

Universiteit Utrecht, Dept. of Information and Computing Sciences,
3508 TB Utrecht, The Netherlands

Abstract. This work proposes a human interaction recognition based approach to video indexing that represents a video by showing when and with whom was interacted throughout the video. In order to visualize the length of an interaction, it is required to recognize individuals that have been detected in earlier parts of the video. To solve this problem, an approach to photo-clustering is extended to video material by tracking detected faces and using the information from tracking to improve the recognition of human beings. The results of the tracking based approach show a considerable decrease of false cluster assignments compared to the original method. Further, it is demonstrated that the proposed method is able to correctly recognize the appearance of five out of the six individuals correctly.

Key words: computer vision, video indexing, bag of words, human recognition

1 Introduction

The current advance of head mountable video recording devices calls for supportive technology regarding the collection of recorded material. The ever increasing size of personal video collections can be approached by an automated video indexing system that supports users by giving concise summaries of lengthy videos. This work proposes a human interaction recognition based approach to video indexing that represents a video by showing when and with whom was interacted throughout the video. By interaction we mean any interaction between the filmmaker and a person that is visible in the video, which includes social interactions like meeting friends as well as non-social interactions like paying the groceries at the store. In Figure 1, several frames are presented that show the cashier of a store interacting with the filmmaker who is paying groceries.

The index we propose is visualized as a horizontal axis representing the time in the video. Each detected interaction partner is represented as a bar above the axis together with a picture extracted from the video. The length and position of the bar represents the moment in time in the video. For an example of an index please see Figure 3, which shows the index representing the ground truth of the dataset we have recorded for the evaluation of this work. We approach the problem of creating a human interaction based index by detecting all individuals using face detection. All detections are clustered based on visual appearance. Optimally, each cluster contains all detections from one individual so that the

first and the last frame within one cluster represents the start and the end time of an interaction.

The aim of our work is to improve the quality of the clustering as well as the visual index by extending the work by Song and Leung (2006) [1] to include video specific information. We identify groups of the same individual by tracking faces through several frames.

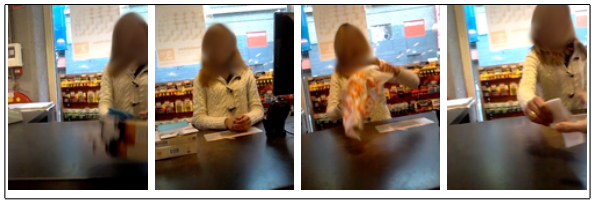

Fig. 1. An example of an interaction is the interaction of the movie maker with a cashier when paying groceries at a shop.

2 Related Work

In the work by Song and Leung (2006) [1], consumer photo albums are clustered based on the individuals they show. For the detection of individuals face detection is used. The recognition is based on a combined distance measure that considers the appearance of detected faces as well as the appearance of clothes. Clothes are detected by selecting a predefined area below any detected face. To prevent the clothes detections of two individuals that are near to each other from overlapping with the other individual's clothes, the clothes detections are segmented by maximizing the difference between the color histograms of both detections. We propose an extension to the work by Song and Leung (2006) [1] to adapt to egocentric video material.

Everingham, Sivic and Zisserman (2009) [2] propose a similar approach of integrating information from tracking into a distance measure between two detection groups. Opposed to our approach, they use the maximum distance between all possible pairs of detections within the detection groups as the distance between two detection groups. In our work, we build a single descriptor per detection group, modeling appearances and its changes as a statistical model. By including information from multiple detections into the descriptor, we improve the quality of the measure of distance between two detection groups that are known to contain detections of the same individual.

3 Approach

We assume that any interaction between the filmmaker and another person involves looking at each other, meaning an interaction can be detected by repeat-

edly recognizing a frontal human face. Thus, to detect interactions, we apply face detection as proposed by Viola and Jones (2001) [3], which detects human beings looking at the camera. We track faces using optical flow as described by Everingham et al. (2009) [2], resulting in detection groups that are known to contain detections of the same individual. To prevent bypassing people from being shown as interaction partners, the detection groups can be required to contain a minimum number of detections.

Optical flow cannot always track a face during the complete interaction, for example because the interaction partner is temporarily out of sight or occluded. That means that an interaction can consist of one or more detection groups. In order to recognize when an interaction starts and where it ends, all detection groups belonging to the same interaction need to be found. We approach this problem by clustering all detections based on their visual appearance into a set of clusters. From the clusters we learn the first and the last frame of the interaction.

For the process of clustering we use and extend the approach of Song and Leung (2006) [1], which combines face recognition with clothes recognition to cluster consumer photo sets. In this section, we first give a summary of how the original approach works and then present the extensions that we propose in order to improve the clustering results when applied to egocentric video material.

In the approach of Song and Leung (2006) [1], for each detection of clothes, overlapping patches are extracted to form a global bag of patches. In order to prevent parts of skin that might overlap clothes, a skin detector is trained based on extracted skin patches from below the eyes. Patches that are classified as skin are ignored in the further. To all patches from all detections principal component analysis is applied. The first 15 principal components are clustered by k-means to form a dictionary of visual words. The patches of each detection are quantized with respect to the visual word dictionary and each bin is multiplied with $\log(\frac{1}{w_i})$, with w_i being the fraction of all patches that has been assigned to the ith bin of all descriptors. This adjustment emphasizes rare patches as they contain more information compared to patches that are frequent in all detections. The dot product of both vectors gives the distance between two clothes detections. Regarding the distance between two faces, we use descriptors based on facial features as described by Everingham et al. (2009) [2]. The distance between two face descriptors is given by the Euclidean distance between the two vectors.

To combine the two distance measures between faces and between clothes into a single distance value between two detections, Song and Leung (2006) [1] uses linear regression. The probability that two detections are the same person is given by

$$P(Y = 1|x_f, x_c) = \frac{1}{1 + exp(-w_f x_f - w_c x_c - w_0)}, \tag{1}$$

where x_f is the similarity between the faces and x_c is the similarity between clothes. The weights w_f and w_c control how much influence each of the similarities has on the combined outcome, and w_0 provides an offset. Given a labeled

training set, the values for w_0, w_c and w_f can be learned by applying iterative reweighted least squares. In this work the values has been chosen experimentally due to the lack of an appropriate data set.

Based on the combined distance measure, the affinity matrix is calculated, holding all pairwise similarities between all detections. The similarity matrix is used to apply spectral clustering, with the clusters being the desired grouping of the detected individuals. We currently set the number of clusters manually, but existing heuristics to estimate the correct number of clusters can easily be implemented. The work of Luxburg (2007) [4] presents several approaches to this problem.

Until here, the approach of Song and Leung (2006) [1] is explained and will be referred to as the original approach in our experimental evaluation. In the following, we propose two extensions, multiple dictionaries of visual words and descriptors for multiple detections.

3.1 Multiple Dictionaries of Visual Words

Using multiple dictionaries has the advantage that the dictionaries can be calculated before the data is completed, which is important for possible real time applications. In our application, detection groups are a good candidate for having an individual bag of words dictionary. When using separate dictionaries per detection group, it is not possible to compare descriptors to each other directly as they refer to different dictionaries. In the following, two different approaches are presented that can solve this issue by integrating the dictionaries into the distance measure between two descriptors.

Multiple separate Dictionaries Aly, Munich and Perona (2011) [5] propose to approach multiple dictionaries by building a descriptor per detection per dictionary. We evaluate this approach in our experiments.

Multiple dictionaries with the Earth Movers Distance Another possibility is to change the distance measure used to compare the descriptors. Instead of using the dot product between two descriptors, the Earth Mover's Distance (EMD) can be used, which allows to take into account the distance between the dictionaries as well. The EMD can be though of the minimum amount of work that is required to fill holes in the ground with earth from piles in some distance to the holes.

In [6], the EMD between two signatures $P = \{(p_1, w_{p1}), \ldots, (p_n, w_{pn})\}$ and $Q = \{(q_1, w_{q1}), \ldots, (q_n, w_{qn})\}$ is defined as

$$\mathbf{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}, \tag{2}$$

where $\mathbf{D} = [d_{ij}]$ is the ground distance matrix, with d_{ij} holding the distance between p_i and q_j, and $\mathbf{F} = [f_{ij}]$ being the flow matrix, holding the flows between weights w_{pi} and w_{qj} that minimize the overall cost

$$WORK(P, Q, \mathbf{F}) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}, \tag{3}$$

subject to

$$\begin{aligned} f_{ij} &\geq 0 && 1 \leq i \leq m, 1 \leq j \leq n \\ \sum_{j=1}^{n} f_{ij} &\leq w_{pi} && 1 \leq i \leq m \\ \sum_{i=1}^{m} f_{ij} &\leq w_{qj} && 1 \leq j \leq n \\ \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} &= \min(\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj}). \end{aligned} \tag{4}$$

Intuitively, applying the EMD to compare descriptors of visual words can be explained as follows: When comparing two descriptors, the visual words of the dictionaries they refer to represent the location of the piles and holes that are to be filled. The visual words of the first descriptor represent the location of the piles and the visual words of the second descriptor represent the location of the holes that are to be filled. The distance between two visual words is the difference in their appearance. As the descriptors are histograms of visual words, their entries represent the size of the holes and the weight of the piles. Consequently, high values in the histograms can only be *moved* to the other descriptor at low cost when the visual words they represent have a similar appearance and similar size.

Formally speaking, the first part of the signature, $p_1, \ldots, p_n$, are the back-projected visual words. The visual words are back-projected because each *detection group* uses an individual PCA transformation vector. As a result, the principal components cannot be compared with each other in PCA space, because the principal components of each group have a completely different meaning in original space. The second part of the signature, the weights $w_1, \ldots, w_n$, are the histograms of visual words. The ground distance matrix $\mathbf{F}$ is calculated by taking the L1-distance between every two visual word vectors p_i and q_j. By normalizing the histograms to sum up to 1, all descriptors have the same amount of weight and the EMD between them becomes a metric.

3.2 Descriptors from multiple Detections

In the original approach by Song and Leung (2006) [1], the appearance of each detection is modeled as a histogram of visual word frequencies. To build a descriptor that contains information of more than one of such descriptors, this work evaluates a way of modeling multiple histograms as a single descriptor.

The most simple approach is to build a vector containing the average values of each bin. The advantage of this approach is that the effect of outlining values is reduced significantly. On the other hand, bins with a high variance within one

detection group will not be represented appropriately, as the information about the variance is lost.

A better approach is to use two vectors containing not only the average, but also including the standard deviation for each histogram bin. In histograms or a vector of average bin values, the values of each bin cannot be compared by simply subtracting the values. Instead, each bin is interpreted as a Gaussian distribution and the distance between two bins is the probability of them describing each other. For this, the normalized L2 distance between two Gaussian distributions is used, which is described in the work of Jensen, Ellis and Christensen [7] as

$$d_{nL2}(p_1, p_2) = \int (p_1'(x) - p_2'(x))^2 \mathrm{d}x, \tag{5}$$

where

$$p_i' = p_i(x) / \sqrt{\int p_i(x)^2 \mathrm{d}x}. \tag{6}$$

In the implementation, $\int p_i(x)^2 \mathrm{d}x$ is approximated by sampling 1000 linear data points between 0 and 1.

4 Experiments

For the experimental evaluation of the proposed extensions we recorded a custom dataset. The dataset consists of recordings of paying groceries from an egocentric perspective and shows 6 different individuals. The resulting number of detections and detection groups resulting from applying face detection and tracking the detected faces with optical flow is shown in Table 1. The face detector is the only component that requires training, but since we use the trained classifier from OpenCV we have no training phase at all and consequently use the complete dataset for testing.

Individual	#1	#2	#3	#4	#5	#6
Detections	59	201	52	44	81	66
Detection Groups	3	3	3	2	3	5

Table 1. Number of detections and detection groups in the evaluation dataset from face detection and tracking.

4.1 Clustering performance

In order to evaluate the proposed extensions we first measure the resulting receiver operating characteristics (ROC) of the clustering step. As proposed by Song and Leung (2006) [1], the Rand index as proposed by Rand (1971) [8] is used to calculate the true positive rate (TPR) and false positive rate (FPR): Given a set of N detections,

"[...] any clustering result can be seen as a collection of $N(N-1)/2$ pairwise decision. A false alarm happens when a pair actually from different individuals, but the algorithm claims they are the same individual. A true positive (detection) is when a pair actually from the same individual and the algorithm also claims so" ([1]).

The results of clustering the dataset using the multiple dictionaries bag of words approach with Earth Mover's Distance (MDBOW EMD) as well as the multiple dictionaries bag of words approach with separate dictionaries (MDBOW SD) are compared to the results of the original approach in Table 2. The number of clusters C is initially set to $C = 2$ and increased with increasing step size to calculate the receiver operating characteristic. The correct number of clusters is $C = 6$. For this experiment, only descriptors from clothes are used because the approaches do not influence the distance measure between detected faces.

For the correct number of clusters $C = 6$, the MDBOW EMD approach has a 21% lower FPR than the original approach, while the TPR only decreases about 4%. When the detections are clustered into $C = 12$ clusters, the FPR of the same approach is 62% higher, which show the importance of having the correct number of clusters. When using the MDBOW SD, compared to the original approach the TPR is decreased and the FPR is increased independent of the number of clusters that has been set.

Approach	C=2		C=4		**C=6**		C=12		C=18	
	T	**F**	**T**	**F**	**T**	**F**	**T**	**F**	**T**	**F**
Original	1.00	0.73	0.99	0.28	0.98	0.28	0.98	0.29	0.97	0.29
MDBOW EMD	0.99	0.48	0.95	0.23	0.94	0.22	0.98	0.47	0.92	0.22
MDBOW SD	0.97	0.46	0.97	0.45	0.96	0.45	0.96	0.44	0.95	0.44

Table 2. Measuring clustering performance: The resulting true positive rate (T) and false positive rate (F) for the original approach and the two multiple dictionary bag of word approaches using the Earth Mover's Distance (MDBOW EMD) and separate dictionaries (MDBOW SD) as described in Section 3.1. The number of ground truth clusters is $C = 6$.

In Table 3, the results of clustering the dataset with detection group based descriptors are compared to those of the original approach. Again, the number of clusters C is initially set to $C = 2$ and increased with increasing step size to calculate the receiver operating characteristic. The correct number of clusters is $C = 6$. For this experiment, both face and clothes descriptors are used as the proposed method applies to both. To relate the measurement to the preceding experiment, the true and false positive rate are calculated based on the detections and not the detection-groups.

For the correct number of clusters $C = 6$, clustering detection groups results in a TPR decreased by 1% and a FPR decreased by 89%. When increasing the number of cluster to 10, which is 1.6 times the correct number of clusters, the FPR becomes 0.001 and the TPR is 0.95.

Approach	C=2		C=4		**C=6**		C=8		C=10	
	T	**F**	**T**	**F**	**T**	**F**	**T**	**F**	**T**	**F**
Original	1.00	0.73	0.99	0.28	0.98	0.28	0.98	0.29	0.97	0.28
DG	1.00	0.37	1.00	0.09	0.97	0.03	0.96	0.02	0.95	0.001

Table 3. Measuring clustering performance: The resulting true positive rate (T) and false positive rate (F) for the original approach and the detection group based descriptors (DG) as described in Section 3.2. The number of ground truth clusters is $C = 6$.

4.2 Retrieval performance

To analyze the quality of the different bag of words based visual descriptors, the average precision of the first K neighbors is looked at. Since all detections from the same detection group are already known to belong to the same cluster, only those detections are considered which are in different detection groups than the original detection, when retrieving the nearest neighbors. Given a set of detections D and the K nearest neighbors of each detection K_D, the average precision p is the number of correct neighbors divided by the total numbers of neighbors, given by

$$p = \frac{\sum_{d \in D} \sum_{k \in K_D} knn(d,k)}{\sum_D \sum_{K_D} 1} \quad \text{, where} \quad knn(d,k) = \begin{cases} 1 & \text{if same person} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The average precision of retrieving the first 10 nearest neighbors for each detection from our complete dataset is shown in Figure 2. The precision of the original approach for $K = 1$ is 0.82 and decreases approximately linearly to 0.68 for $K = 10$. The MDBOW EMD approach and the MDBOW SD approach have an overall better precision, which is about the same for both approaches. For $K = 1$ the precision for the two new approaches is about 0.95 and decreases to 0.8, which is about 18% better compared to the original approach.

4.3 Resulting index

The ground truth index for our dataset is depicted in Figure 3, whereas ground truth refers to the ground truth of the clustering results. Each horizontal rectangle represents the presence of the depicted individual within the video.

In Figure 4, the index is shown that is constructed by the framework. Individuals 1,2 and 5 are detected at the correct position and with approximately the correct number of frames. Individual 6 is detected at the correct position, although clearly several occurrences are not recognized, letting the index entry begin later and end earlier than the entry in the index representing ground truth. For individuals 3 and 4, two separate index entries are shown. While individual 4 is represented correctly, the 3rd individual's last frame is wrongly recognized at the end of the video. Since our algorithm uses the first and the last frame of a cluster to calculate the beginning and the end of an interaction, the bar for individual 3 spans the second half of the index.

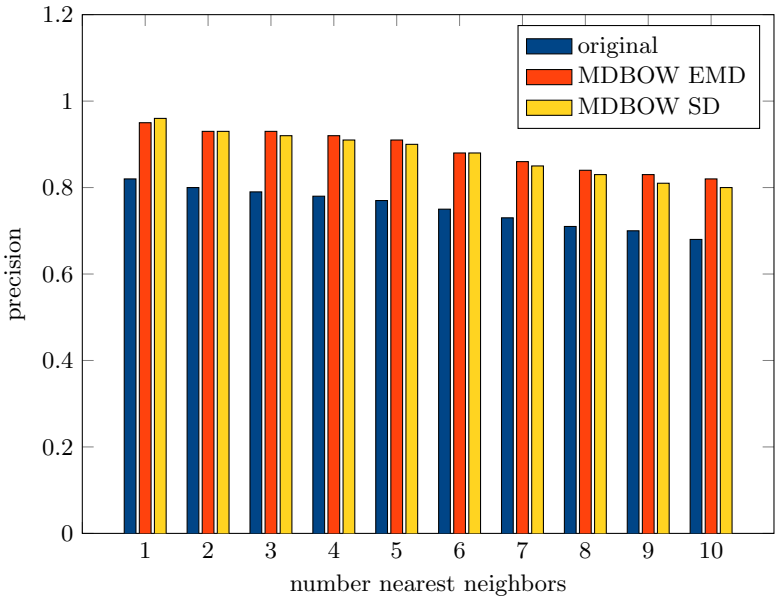

Fig. 2. Measuring retrieval performance: The average precision of different clothes descriptors when retrieving the K nearest neighbors for each detection in dataset 2.

Fig. 3. The index produced for ground truth clustering results.

5 Conclusion

We have presented two different approaches to extend the work by Song and Leung (2006) [1] to cluster egocentric video material: Using multiple dictionaries of bag of words and aggregating separate detections into group-descriptors by using tracking.

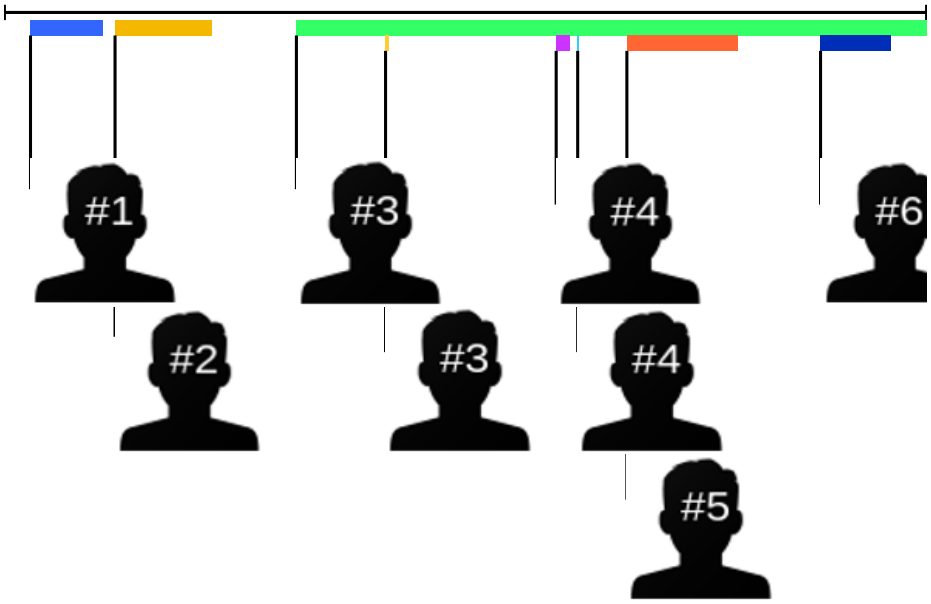

Fig. 4. The index produced for clustering results when using detection group based descriptors.

Regarding the first approach, the experiments show that using multiple dictionaries with the Earth Mover's Distance results in a 21% lower false positive rate than the original approach, while the true positive rate only decreases 4%. Furthermore, the k-NN experiment suggests that both methods perform better than the original approach when used to query for similar detections. In other words, the methods might render very useful in other applications.

The second approach, including multiple detections in a single model, shows good results in our experiments. The false positive rate is reduced by 89%, given the correct number of clusters. Furthermore, we show that the clustering approach can be used to create a time based index representing human interactions for short length video.

References

1. Yang Song and Thomas Leung. Context-aided human recognition - clustering. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision - ECCV 2006*, volume 3953 of *Lecture Notes in Computer Science*, pages 382–395. Springer Berlin Heidelberg, 2006.
2. M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5), 2009.
3. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features, 2001.

4. Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.
5. Mohamed Aly, Mario Munich, and Pietro Perona. Multiple Dictionaries for Bag of Words Large Scale Image Search. In *International Conference on Image Processing (ICIP), Brussels, Belgium, September 2011*, September 2011.
6. Yossi Rubner, Carlo Tomasi, and LeonidasJ. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.
7. Jesper Hojvang Jensen, Daniel PW Ellis, Mads G Christensen, and Soren Holdt Jensen. Evaluation distance measures between gaussian mixture models of mfccs. In *ISMIR 2007: Proceedings of the 8th International Conference on Music Information Retrieval: September 23-27, 2007, Vienna, Austria*, pages 107–108. Austrian Computer Society, 2007.
8. W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.