Space-time Histograms And Their Application To Person Re-identification In TV Shows

Rémi Auguste, Jean Martinet, Pierre Tirilly

Space-time Histograms And Their Application To Person Re-identification In TV Shows

Rémi AUGUSTE
remi.auguste@univ-lille1.fr

Jean MARTINET
jean.martinet@univ-lille1.fr

Pierre TIRILLY
pierre.tirilly@univ-lille1.fr

CRIStAL (UMR CNRS 9189) – Lille 1 University – Villeneuve d'Ascq, France

ABSTRACT

The annotation of video streams by automatic content analysis is a growing field of research. The possibility of recognising persons appearing in TV shows allows to automatically structure ever-growing video archives. We propose a new descriptor to re-identify persons featured in videos, that is to say, to spot all occurrences of persons throughout a video. Our approach is dynamic as it benefits from motion information contained in videos, whereas the static approaches are solely based on still images. We extract persontracks from videos and match them using a new descriptor and its associated similarity measure: the space-time histogram. The originality of our approach is the integration of temporal data into the descriptor. Experiments show that it provides a better estimation of the similarity between persontracks. Our contribution has been evaluated using a corpus of real life french TV shows broadcasted on BFMTV and LCP TV channels and on some annotated episodes from "Buffy: the Vampire Slayer". Experimental results show that our approach significantly improves the precision of the re-identification process thanks to the use of the temporal dimension.

Categories and Subject Descriptors

I.4.8 [**Image processing and computer vision**]: Scene Analysis—*Tracking*; D.2.8 [**Software Engineering**]: Metrics—*Performance measures; Complexity measures*

Keywords

person re-identification; video; spatial-temporal descriptor

1. INTRODUCTION

When automatically annotating a large video database, the computational cost and the precision are the main concerns in choosing the most suited algorithm. Usually, one has to find a trade-off between these two constraints depending on one's needs and available resources. One of the main aspects of the video to be annotated is the persons featured in it. We therefore propose a new descriptor to discriminate among the persons in videos, using color, space and time information: space-time histograms. Our approach is lightweight, comparable to color histograms, and yet conveys better re-identification rate.

The goal of re-identification is to assign a unique label to persons' occurrences from a video. It is not to find the name of a person since no identity is involved.

First, we formally define the various notions used in our work. We define a video V_i, part of a corpus of videos V, as a set of ordered shots s_j^i:

$$(V_i, \leq) = \{s_0^i, s_1^i, \ldots, s_{|V_i|-1}^i\} \tag{1}$$

A shot s_j^i of a video V_i is defined as an ordered set of contiguous frames recorded by a single camera:

$$(s_j^i, \leq) = \{f_t, f_{t+1}, \ldots, f_{t+|s_j^i|-1}\} \tag{2}$$

where t is the temporal index of the first frame of the shot of the video. This index induces an ordering relationship ($\leq$) between the frames of a shot. The shot ends when the camera changes or when a cut is encountered. Those two notions can be found in most of the video-based person recognition systems [28].

The region of pixels that compose a person in a frame forms a blob. The sequence of sucessive blobs of a single person in a shot form a persontrack [7]. In order to extract persontracks from a video shot, the consecutive detections are merged using a tracking algorithm. We define the unordered set of persontracks $\mathbb{O}_i$ from a video V_i as:

$$\mathbb{O}_i = \{o_0, o_1, \ldots, o_{|\mathbb{O}_i|-1}\} \tag{3}$$

We define the set of identities featured in all videos as:

$$\mathbb{I} = \{\iota_0, \iota_1, \ldots, \iota_{|\mathbb{I}|-1}\} \tag{4}$$

and the ground truth function id that associates a persontrack o to its identity ι:

$$\begin{aligned} id: \quad & \mathbb{O} \to \mathbb{I} \\ & o \to \iota \end{aligned} \tag{5}$$

The objective of person re-identification is to label the persontracks using a unique label $\Omega_{\iota,i}^*$ for each identity. The result is a set of labeled persontracks:

$$\Omega_{\iota,i}^* = \{o \in \mathbb{O}_i | id(o) = \iota\} \tag{6}$$

Our objective is to propose a discriminative signature for persontracks. This signature is used to select a label for each track. The proposed signature is dynamic as it uses temporal features of the videos. Most approaches of the state of the art do not consider time in their description of a persontrack [3]; they only provide a representation based on the static visual aspect of the persons.

ICMR'15, June 23–26, 2015, Shanghai, China.

ACM 978-1-4503-3274-3/15/06$15.00.
http://dx.doi.org/10.1145/2671188.2749332.

2. RELATED WORK

A typical re-identification approach is global, as it is based on the overall appearance of the persons (e.g. the upper-body, the complete body, the silhouette, etc.) to compute a signature. Local approaches use interest points to compute the signatures. Those approaches require for that a higher image quality and pose numerous constraints concerning the pose or illumination of the persons. Furthermore, local approaches usually require a classification algorithm to re-identify the persons. The classification is trained on annotated data; this goes against the fully unsupervised aspect of re-identification. For these reasons, local approaches are seldom used in the context of person re-identification [3].

Several methods for person re-identification can be found in the litterature. Bird et al.[5] presented a multi-camera re-identification system where pedestrians are detected and annotated. They are re-identified based on the color of their clothes and the way these colored region are correlated. A Linear Discriminant Analysis (LDA) is applied to accentuate the difference between the individuals in the feature space. It simplifies the matching of a person going from one camera to another. In a multi-camera person re-identification context, Fu et al. [10] proposed an image-based approach robust to some degrees of corruption. This is done by dividing each image of a person into color stripes. The different views of the person are used to detect corrupted area of a view and estimating the missing or corrupted data from other views. Their approach require training data for each person in order to re-identify them. The approach of Ngo et al.[18] is based on tracking interest points found on faces using the algorithm of Shi et al.[22]. In order to track such interest points, the authors apply an optical flow algorithm and count the number of points shared by both faces. Above a given threshold, the faces are labeled as belonging to the same person. In a similar way, Hamdoun et al.[14] use interest points obtained by a method inspired by SURF[2]. In their work, the signature is built from the interest points of a persontrack. The authors calculate the sum of absolute differences (SAD) between the set of points of the known persontrack and the set of the query persontrack for matching purposes. Zheng et al.[29] formulated person re-identification as a distance learning problem. The aim is to learn an optimal distance that maximises the matching accuracy regardless of the choice of representation. They introduced a probabilistic model that learns this distance from correct and incorrect examples. Recently, Gandhi and Ronfard[12] re-identified the characters from Alfred Hitchcock's movie "Rope" (1948) by using an appearance model based on colored ellipses. This approach does not use a person detector; instead, a sliding window moves across each frame and generates an appearance model. If the model matches one of the characters, the position of the person is saved. This approach allows person re-identification in many cases, with fair robustness to occlusions. The main drawback is that the method generates many false detections. In a similar way, Zeng et al.[27] proposed to model the persons by a color topology. This representation takes into account how color regions are connected one to another. Their approach gives good results for pedestrian re-identification. The main issue of this approach is that the person have to be featured in the same way each time in order to be re-identified correctly.

Several approaches are based on histograms to re-identify the persons[24, 21, 15]. In the work by Truong Cong et al.[24], passengers of a moving train are re-identified across two cameras positioned in a single coach. For each person, a color histogram, a spatiogram and a color-path are built separatly and avaluated. The spatiograms were defined by Elmongui et al.[8], who proposed to use them for classification and search in numerical databases. The performances of the various approaches were evaluated and the color histograms did not obtain good re-identification scores. Spatiograms obtained the best results, which were slightly better than those of color-path. In the work of Schwartz et al.[21] et Hirzer et al.[15] color histograms obtained good re-identification results for face identification against Local Binary Patterns (LBP) and Histograms of Oriented Gradients (HOG).

3. SPACE-TIME HISTOGRAMS

To benefit from color, geometry and motion information from videos and to distinguish persons better, we extended the spatiogram to the temporal aspect and applied them to videos. The proposed space-time histograms are lightweight, and are easy to compute and to compare.

3.1 Definition

Space-time histograms are an extension of the spatiograms proposed in [4], which are themselves an extension of the classic color histograms. The data structure of the space-time histogram sth_o built on a persontrack o is defined as:

$$\mathrm{sth}_o(b) =< n_b, \mu_b, \Sigma_b >, \quad b = 1, \ldots, B \tag{7}$$

where n_b is the number of pixels in bin b and B is the total number of bins. The average position in space and time, μ_b, is defined as:

$$\mu_b = (\bar{x}_b, \bar{y}_b, \bar{t}_b) \tag{8}$$

where $\bar{x}_b$, $\bar{y}_b$ and $\bar{t}_b$ are the average normalized positions of the pixels in space and time. Σ_b is the covariance matrix of the space-time positions:

$$\Sigma_b = \begin{pmatrix} cov(x_b, x_b) & cov(x_b, y_b) & cov(x_b, t_b) \\ cov(y_b, x_b) & cov(y_b, y_b) & cov(y_b, t_b) \\ cov(t_b, x_b) & cov(t_b, y_b) & cov(t_b, t_b) \end{pmatrix} \tag{9}$$

This covariance matrix is symetric since $cov(a, b) = cov(b, a)$. Space-time histograms, as defined in Equation 7, contain spatiograms, in the same way, spatiograms contain the color histograms. In the terminology of [4], space-time histograms could be called third order spatio-tempo-gram. The complexity of space-time histograms is comparable to that of other approaches that consider the temporal data of videos like cumulative color histograms or cumulative spatiograms.

3.2 Time complexity

The calculatory cost $T_{\text{cumul}}(p, f, B)$ of the construction of space-time histograms, cumulative spatiograms and color histograms can be estimated by the following formula:

$$T_{\text{cumul}}(f, p, B) = O(f \times p + B) \tag{10}$$

where p is the number of pixels per frame, f the number of frames and B the number of bins of the considered histogram.

Another approach, where each frame from a video sequence could be represented independently by several spatiograms or several color histograms is possible. In this case, the calculatory cost of the construction is higher since each frame is associated to a descriptor. The final step has to be applied to each descriptor. The construction cost $T_{\text{ind}}(p, f, B)$ in such case is estimated by the formula:

$$T_{\text{ind}}(p, f, B) = O(f \times (p + B)) \tag{11}$$

This calculatory cost does not take into account the cost of allocating a new descriptor containing B bins; however, it is necessary to allocate as many descriptors as there are frames in the video. When the video duration is more than a few seconds, the allocation cost becomes non-negligible over the total cost of the construction.

3.3 Space complexity

Let d be the position number of dimensions of the position of the pixels considered. Color histograms do not take into account any position, thus $d = 0$ in this case. Spatiograms take only into account spatial data so $d = 2$ in this case. Finally, space-time histograms consider the space-time position of the pixels, $d = 3$ in this case. This allows us to express the memory cost $M_{\text{cumul}}(B, d)$ of space-time histograms, cumulative spatiograms and cumulative histograms as a function of d:

$$\begin{aligned} M_{\text{cumul}}(B,d) &= O(c \times B), c = 1 + d + \frac{d(d+1)}{2} \\ &= O(d^2 \times B) \end{aligned} \quad (12)$$

where B is the number of bins and c the amount of data stored by the model. For example, in the case of space-time histograms, $c = 10$ because: the pixel count has a cost of 1, the average position $\bar{x_b}, \bar{y_b}, \bar{t_b}$ has a cost of 3 and the covariance matrix has a cost of 6. The covariance matrix is of dimension 3×3, but because of its symmetry, only 6 elements need to be memorised.

In the case where the descriptor is built independently for each frame, the memory cost $M_{\text{ind}}(B, d, f)$ is again much higher than the one cumulating over the frames:

$$\begin{aligned} M_{\text{ind}}(B,d,f) &= O(f \times c \times B) \\ &= O(f \times d^2 \times B) \end{aligned} \quad (13)$$

3.4 Similarity measure

In order to compare space-time histograms, we propose a similarity measure inspired from the measure used for spatiograms [24]. A temporal dimension was added to it. Space-time histograms combine the frequencies of colors with space-time distributions of pixels.

To measure whether two histograms come from the same statistical distribution, the Mahalanobis distance is used. Let ψ_b be the measure based on the Mahalanobis distance, measuring the similarity of the bins of index b coming from two space-time histograms.

$$\psi_b = 1 - \sqrt{(\mu_b - \mu'_b)^t \hat{\Sigma}_b^{-1} (\mu_b - \mu'_b)} \quad (14)$$

In Equation 14, the covariance matrix is estimated using the following formula:

$$\hat{\Sigma}_b^{-1} = (\Sigma_b^{-1} + (\Sigma'_b)^{-1}) \quad (15)$$

The χ^2 distance has the property of measuring the dissimilarity between two bins that is proportionnate to their size. This property is interesting in that a large bin with a small difference between two space-time histograms will not influence the measure too much.

We have defined χ_b^2 as similarity measure between two bins of index b as:

$$\chi_b^2(n_b, n'_b) = 1 - \frac{(n_b - n'_b)^2}{n_b + n'_b} \quad (16)$$

The combination of this measure with Mahalanobis' measure allows us to take into account the various aspects of the space-time histograms. This similarity measured between two space-time histograms sth_o and $sth_{o'}$ of identical size is defined as:

$$s(sth_o, sth_{o'}) = \sum_{b=1}^{B} \psi_b \times \chi_b^2(n_b, n'_b) \quad (17)$$

3.5 Comparison complexity

The calculatory cost of space-time histogram comparison $T_{\text{sim\_cumul}}(B, d)$ is:

$$T_{\text{sim\_cumul}}(B,d) = O(B \times d^3) \quad (18)$$

This cost is similar for spatiograms and color histograms built cumulatively.

The descriptor built for individual frames cannot be compared using a similar algorithm like cumulatively built descriptors. This is due to the possible length difference between the video sequences. Dynamic Time Warping can solve this issue of *n-by-m* comparison between two timely ordered sequences of descriptors (or observations). This algorithm usually has a complexity of $O(f^2)$, that can be improved in numerous ways [26]. The total comparison cost $T_{\text{sim\_ind}}(f, B, d)$ is:

$$T_{\text{sim\_ind}}(f,B,d) = O(f^2 \times B \times d^3) \quad (19)$$

The descriptors built cumulatively upon every frames of the video have much lower time and complexities, than those considering independently every frame.

4. PERSON RE-IDENTIFICATION

The proposed similarity measure between two space-time histograms can be used on video-based persontracks to re-identify persons. Our hypothesis is that the similarity $s(\text{sth}_o, \text{sth}_{o'})$ is high (close to 1) when $id(o) = id(o')$, in other words, when two persontracks contain the same person. On the contrary, this similarity is low (close to 0) when $id(o) \neq id(o')$. This hypothesis requires that the compared persontracks are taken from the same video V_i, so that the global appearance of the persons shows minor variations.

Once the persontracks have been described as space-time histograms, they can be compared by measuring their similarity. A similarity matrix M is generated to keep the comparison results. This matrix will be used to measure the precision of our approach. Because our similarity measure is symmetric, the matrix is also symmetric.

The first step of our approach is to build a space-time histogram for each persontrack and then calculate the similarity matrix.

This matrix can be seen as a set of lines:

$$M = \left[M_1, \ldots, M_{|M|}\right]^T \quad (20)$$

where each row M_i of the matrix M gives the similarity measure between a space-time histogram sth_{o_i} and every space-time histograms.

Using the lines of the similarity matrix M, we sort each one in descending order of similarity. We define a matrix R containing, in each row, the sorted similarities of each persontrack taken from matrix M:

$$R = \{r_{ij} = o_k | \, \text{rank}(o_k, M_i) = j\} \quad (21)$$

where $\text{rank}(o_k, M_i)$ is the rank of the value of the similarity S_{ik} of the persontrack o_k in the line M_i. When we consider this rank in descending order of similarity: rank 1 for the highest similarity, rank 2 for the second higher, etc. The first value is a self-match: the similarity measure between the space-time histogram representing the persontrack and itself (value 1).

4.1 Evaluation Metric

For each line M_i, we want to measure the matching's precision between o_i and the other persontracks in the video bearing the same identity as o_i. The problem of the classic precision measure is that the number of persontrack per identity may vary. The identities

with a few number of persontracks would contribute too much to the average without this. We used the *precision at n* (*P@N*) [19]. It is the precision measured by considering the n first elements, where n is the number of correct answers. In our approach, for the persontrack o_i, n_i is the number of persontracks of a video V_v bearing the same identity as o_i:

$$n_i = |id^{-1}(id(o_i))|, \forall o_i \in \mathbb{O}_v \tag{22}$$

This precision at n_i for o_i is given by P_i:

$$P_i = \frac{|\{r_{ij} \in R | j \leq n_i\} \cap id^{-1}(id(o_i))|}{n_i} \tag{23}$$

We then calculate the weighted average $\bar{P}$ of the precisions values:

$$\bar{P} = \frac{\sum_{i=1}^{|M|} P_i n_i}{\sum_{i=1}^{|M|} n_i} \tag{24}$$

This average is weighted by the number of persontracks by identity. It avoids the introduction of a bias in the final metric.

4.2 Data

The data we use was distributed in the context of the ANR[1] REPERE challenge[2][11]. It contains several hours of annotated french TV shows from LCP and BFMTV channels. Several shows from these channel are included in this dataset, with large variation in length and settings. Some shows contain outdoor scenes. For our experiments, we manually inspected and filtered the original dataset to remove any ambiguity and to provide a sound groundtruth for video-based person re-idenfication. Our dataset called FoxPersonTracks is available[3] and thoroughly described in [1].

In total, the subset we considered for our experiments consists of 303 different persons whose names are given in the annotations. Each person appears in average in 15 different shows. The anchors have a larger number of persontracks as they are visible more frequently than the other persons. They can appear more thant 50 times per show where some persons can only appear once.

4.3 Data filtering

The persontracks are extracted from 141 videos. We used the annotations to make sure that each persontrack contains only one person. Some faces in the REPERE dataset are not annotated because of size or semantic criteria. Furthermore, the faces position are annotated on the keyframes but the annotated parts do not take into account the shot boundaries. We used the Viola and Jones face detection algorithm[25] to detect the presence of more than one face in a persontrack. Then, we manually filtered each persontrack to ensure the quality of our dataset. This avoids any confusion between the persons during the evaluation.

At the end of the filtering we have 5279 video persontracks of 303 different persons. Each person appears in average in 5 videos. The anchors are more represented than the other persons.

4.4 Persontrack extraction

The persontracks are extracted from the video by first detecting the person in each frame using Viola and Jones face detector[25] (see Figure 2a). The detection is optimised to maximise the proportion of skin color in the detection (see Figure 1). This is done to eliminate most of the background from the detection when the person is not perfectly facing the camera. The calculated face's position and size is used to initialise a set of masks provided to Grabcut algorithm[20] in order to separate the person from the background (see Figure 2c). The masks are initialized from ellipses calculated from the detected face (see Figure 2b). The extracted person's image is then resized to a fixed size. This size was empirically deduced from our experimentations, we noticed that using 50% of the original frame size was good to fit most of the occurrences without resizing them. This is done so the person's position are centered and the number of pixels in each frame is normalized. The illumination is not normalised in our approach as the illumination is stable enough throughout a TV show. This step should otherwise be done.

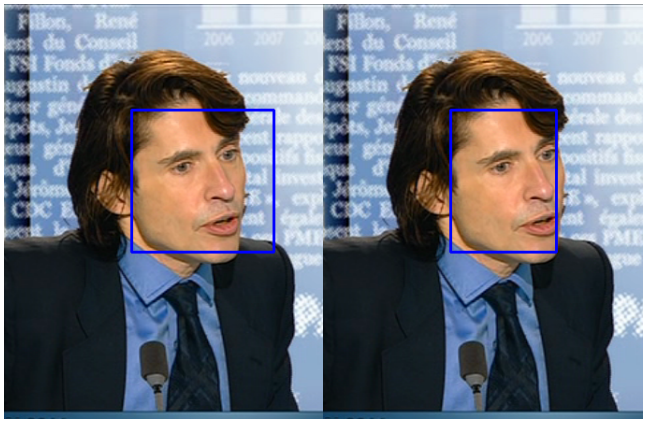

Figure 1: Example of face detection before and after our optimisation based on the skin color proportion.

After the persontrack extraction, we obtain short videos, each centered on one person in front of a zero value pixel background. Our approach is therefore robust to camera zooming or panning. The persontrack can then be viewed as a normalised volume as shown in Figure 3, composed of the extracted person stacked through time. The depth of the volume is the temporal dimension. Space-time histograms, spatiograms and color histograms are built upon those extracted persontracks. During the construction, the background pixels are ignored.

5. EXPERIMENTS

We will now evalutate the space-time histograms and compare its precision to other approaches.

5.1 Precision

We compare the evolution of the precision of different approaches where the number of bins vary. This way, we can study their behavior and compare their precision.

We observe in Figure 4 that the precision of each approach evolves parallelly. The precision of each approach increases rapidly for a low number of bins (between 10 and 1,000). The precision reaches a maximum and starts decreasing at around 5,000 bins. Our approach based on space-time histograms (sth) obtained a better precision than color histograms (h) or spatiograms (sp). This improvement in precision is very significant as the p-value of the student-test is very low (close to zero). This confirms our hypothesis that space-time information is important to re-identify the persons in

[1] Agence Nationale de la Recherche
[2] REPERE Evaluation Package,
ELRA catalog (http://catalog.elra.info),
ISLRN: 360-758-359-485-0, ELRA ID: ELRA-E0044
[3] FoxPersonTracks dataset,
ELRA catalog (http://catalog.elra.info),
ISLRN: 168-132-570-218-1, ELRA ID: ELRA-S0374

(a) Original frame with the detected face.

(b) Mask calculated from the detected face.

(c) Result of the Grabcut algorithm applied on the original image using the mask.

Figure 2: Example of our person extraction process on a single frame and a single person.

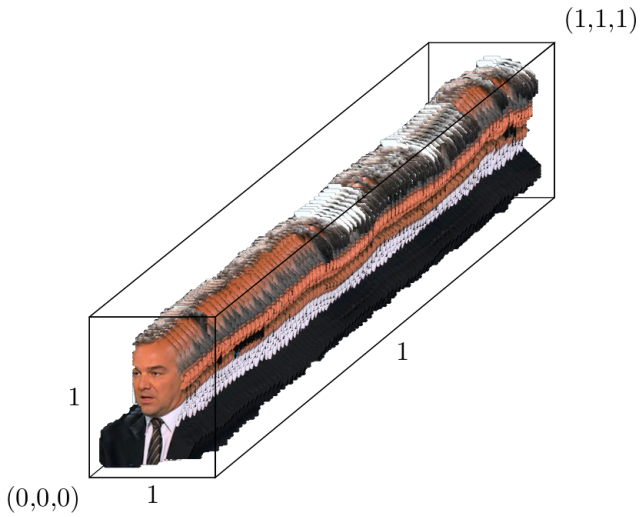

Figure 3: Example of extracted persontrack from a BFMTV TV show.

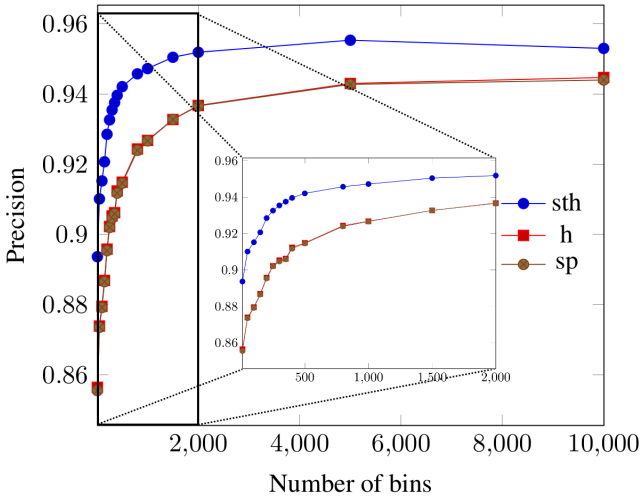

Figure 4: Evolution of the precision as the number of bins in the descriptors increases, from 10 to 10,000 bins. The curves of color histograms and spatiograms are overlaping.

persontracks.

It is interesting to observe that the precision of spatiograms is almost identical to that of color histograms. This indicates that space information along with color is not better at distinguishing between the persons featured in the persontracks than color alone.

5.2 Complexity

In this experiment, we only compare their memory cost. Color histograms has a memory cost of 1 by bin (the data count). Spatiograms have a memory cost of 6 and space-times histograms have a memory cost of 9 (cf. Section 3).

Figure 5 shows the precision of the different approaches with their relative (to color histograms) memory cost relative. We observe that for a memory cost lower than 4,500, color histograms yield the best precision. For a memory cost of 4,500, space-time histograms and the color histograms have a similar precision. This means that a 500 bins space-time histogram is equivalent in precision to a 4,500 bins color histogram. With a memory cost higher than 4,500, space-time histograms give the highest precision that is even increasing, whereas the precision of color histograms decreases slowly. Thus, a color histogram cannot reach the precision of a space-time histogram using more than 500 bins.

Spatiograms give a much lower precision, relative to their memory cost, than the other approaches. It is only with a memory cost of over 30,000 than the precision of the spatiograms reaches and overtakes that of the color histograms. Therefore a 5,000 bins spatiogram is equivalent in precision to a 30,000 bins color histogram. The precisions of spatiograms and space-time histograms seem to evolve parallel one to another. Thus, the precision of spatiograms cannot match the one of space-time histograms.

In conclusion, the space-time histograms yield a higher precision, for an equivalent memory cost, than color histograms and spatiograms. This result clearly shows the contribution of temporal and spatial information to re-identify the persons featured in the persontracks.

In order to compare our approach to other approaches, we evaluated it to the "Buffy" dataset [23]. We used the detections provided in the dataset along with the identities as groundtruth. In order to

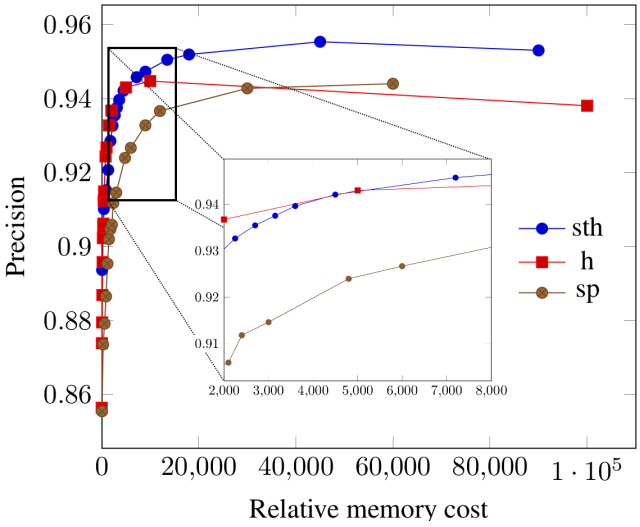

Figure 5: Evolution of the precision as the memory cost increases for each approach.

	Episode 05-02		Episode 05-05	
Approach	$\bar{P}$	# of tracks	$\bar{P}$	# of tracks
STH (All)	0.455	876	0.373	663
STH (Mixed)	0.414	677	0.498	544
STH (Frontal)	0.463	337	0.492	324
Sp (All)	0.333	876	0.370	663
Sp (Mixed)	0.376	677	0.458	544
Sp (Frontal)	0.428	337	0.457	324
H (All)	0.285	876	0.368	663
H (Mixed)	0.328	677	0.409	544
H (Frontal)	0.380	337	0.408	324
[9]	0.682	516	0.692	477
[17]	0.312	491		

Table 1: $\bar{P}$ measure of our approach against other on episode 2 and 5 of the fifth season of "Buffy: the Vampire Slayer".

extract the persontracks, we applied the process presented in Section 4.4. During the extraction the persontracks annotated as false detections were ignored.

The precision $\bar{P}$ was measured on the entire dataset and on two subsets modeled by space-time histograms (1500 bins, RGB). One subsets includes the persontracks containing at least one frontal face. The other is composed of persontracks containing only frontal faces. The results, given in Table 1, show that our approach performs better with frontal faces. This is due to the fact that the persontrack extraction process is not designed for non-frontal facing persons: it tends to incorporate a lot of background pixels, thus adding noise to the signature. The medium re-identification rate can be explained by the fact that the characters featured in the show change clothes several time during an episode. However, these results show that our approach performs better on this dataset than [17]. Our precision is lower than the one obtained by [9] for a recall of 100%. We also observe that spatiograms and color histograms, yield lower results than space-time histograms.

In order to automatically group the persontracks according to their identity, we applied an agglomerative hierarchical clustering on our similarity matrix (see Table 2). We observe that the purity and the fragmentation both improve when using only frontal facing persontracks. Those results are consistent with those (reported in [16]) obtained by [16], [6] and [13] on the other "Buffy" dataset[13]. Unfortunately the dataset [13] and [9] are using different episodes of "Buffy: the Vampire Slayer", so we cannot compare to them. We do observe that the space-time histograms tend to create less and purer clusters.

6. CONCLUSION AND FUTURE WORK

We introduced a new descriptor to re-identify persons featured in videos called space-time histogram. It takes into account color, spatial and temporal data contained in a persontrack to generate a discriminative signature. This signature is used to re-identify each person in a video. The originality of our approach is the integration of temporal data into the descriptor. Experiments show that it provides a better estimation of the similarity between persontracks. Our descriptor is evaluated using the REPERE dataset [11], featuring real life TV shows broadcasted from BFMTV and LCP TV channels, and the well known "Buffy" dataset [23]. Experimental results show that our approach significantly improves the precision as compared to color histograms and spatiograms for a person re-identification task, thanks to the use of the temporal dimension. Compared to state-of-the-art approaches, our results are lower while having a much lower complexity.

In our future research, we want to improve the persontracks' extraction to obtain higher quality persontracks while taking less time. Our experiments show that the direction the persons are facing has a very high impact on the results.

7. REFERENCES

[1] R. Auguste, J. Martinet, and P. Tirilly. Introducing FoxPersonTracks: a benchmark for person re-identification from tv broadcast shows. *Proceedings of the 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2015.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

[3] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[4] S. T. Birchfield and S. Rangarajan. Spatiograms versus histograms for region-based tracking. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1158 – 1163. IEEE, june 2005.

[5] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167–177, June 2005.

[6] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1559–1566, 2011.

[7] X. W. Duy-Dinh Le and S. Satoh. *Encyclopedia of multimedia*, chapter Face Detection, tracking, and recognition for broadcast video, pages 228–238. Springer-Verlag New York Inc, 2008.

[8] H. G. Elmongui, M. F. Mokbel, and W. G. Aref. Spatio-temporal histograms. *Advances in Spatial and Temporal Databases*, pages 19–36, 2005.

[9] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV

	Episode 05-02			Episode 05-05		
Approach	Purity	Fragmentation	# of clusters	Purity	Fragmentation	# of clusters
STH (All)	0.349	23.26	189	0.409	19.23	149
STH (Mixed)	0.378	17.24	140	0.424	14.93	119
STH (Frontal)	0.428	9.90	83	0.416	11.11	81
Sp (All)	0.332	24.45	213	0.382	21.01	174
Sp (Mixed)	0.398	17.63	159	0.423	17.15	146
Sp (Frontal)	0.425	10.01	93	0.406	11.83	91
H (All)	0.323	24.44	215	0.389	20.69	175
H (Mixed)	0.364	17.60	159	0.416	17.01	148
H (Frontal)	0.423	10.00	92	0.404	11.42	89

Table 2: Clustering analysis of our approach on episode 2 and 5 of the fifth season of "Buffy: the Vampire Slayer".

video. In *Proceedings of the British Machine Vision Conference*, 2006.

[10] M.-H. Fu, Y.-C. Wang, and C.-S. Chen. Exploiting low-rank structures from cross-camera images for robust person re-identification. *IEEE International Conference on Image Processing (ICIP)*, pages 2427–2431, 2014.

[11] O. Galibert and J. Kahn. The first official repere evaluation. *First Workshop on Speech, Language and Audio for Multimedia (SLAM 2013)*, 2013.

[12] V. Gandhi and R. Ronfard. Detecting and naming actors in movies using generative appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3706–3713, 2013.

[13] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 498–505, 2009.

[14] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6, 2008.

[15] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In A. Heyden and F. Kahl, editors, *Image Analysis*, volume 6688 of *Lecture Notes in Computer Science*, pages 91–102. Springer Berlin Heidelberg, 2011.

[16] E. Khoury, P. Gay, and J.-M. Odobez. Fusing matching and biometric similarity measures for face diarization in video. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 97–104, 2013.

[17] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof. Learning to recognize faces from videos and weakly related information cues. In *IEEE Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 23–28, 2011.

[18] T. D. Ngo, D.-D. Le, S. Satoh, and D. A. Duong. Robust face track finding in video using tracked points. *IEEE International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, pages 59–64, 2008.

[19] V. Raghavan, P. Bollmann, and G. S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989.

[20] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314, 2004.

[21] W. Schwartz, H. Guo, and L. Davis. A robust and scalable approach to face identification. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *European Conference on Computer Vision (ECCV)*, volume 6316 of *Lecture Notes in Computer Science*, pages 476–489. Springer Berlin / Heidelberg, 2010.

[22] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[23] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1145–1152, 2009.

[24] D. Truong Cong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In P. Foggia, C. Sansone, and M. Vento, editors, *Image Analysis and Processing (ICIAP)*, volume 5716 of *Lecture Notes in Computer Science*, pages 179–189. Springer Berlin / Heidelberg, 2009.

[25] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2002.

[26] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, 2013.

[27] G. Zeng, H.-M. Hu, Y. Geng, and C. Zhang. A person re-identification algorithm based on color topology. *IEEE International Conference on Image Processing (ICIP)*, pages 2447–2451, 2014.

[28] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35(4):399–458, 2003.

[29] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 649–656, June 2011.