Covariance Descriptor Multiple Object Tracking and Re-Identification with Colorspace Evaluation

Andrés Romero, Michèle Gouiffés and Lionel Lacassagne

Institut d'Éléctronique Fondamentale, UMR 8622, Université Paris-Sud XI, Bâtiment 660, rue Noetzlin, Plateau du Moulon, 91400 Orsay

Abstract. This paper addresses the multi-target tracking problem with the help of a matching method where moving objects are detected in each frame, tracked when it is possible and matched by similarity of covariance matrices when difficulties arrive. Three contributions are proposed. First, a compact vector based on color invariants and Local Binary Patterns Variance is compared to more classical features vectors. To accelerate object re-identification, our second proposal is the use of a more efficient arrangement of the covariance matrices. Finally, a multiple-target algorithm with special attention in occlusion handling, merging and separation of the targets is analyzed. Our experiments show the relevance of the method, illustrating the trade-off that has to be made between distinctiveness, invariance and compactness of the features.

1 Introduction

Multiple objects tracking or matching is a classical task required in most surveillance systems. More than being useful for analyzing trajectories and behaviors in a mono-camera context, it is a challenging issue when objects have to be re-detected from a second camera under different set-ups, or at two very different times. The task faces many difficulties such as scale or appearance change, illumination variations or occlusion. Ideally, the representation of the target has to be chosen so as to be invariant and robust to such phenomena. Unfortunately, most color invariants, although robust against lighting changes, can reduce the separability between the targets and can lead to matching ambiguities. In addition, when targets have a non rigid motion or have low textural or structural contents, the gradient or corner-based methods, such as the classical KLT [1] or SIFT [2] are not appropriate.

Kernel-based methods like Mean-shift [3] are usually well adapted to such objects since they rely on a global statistical distribution. The price to pay is a decrease of discriminant power, therefore several attempts have enhanced the method by background subtraction [4], colorspace switch [5] and by using a spatio-colorimetric histogram [6]. Covariance Tracking [7] is an interesting alternative which employs a compact representation of the correlation between spatial and statistical features within the object window. High performances can

be achieved even for low textured objects, since they are represented by a global model. However, the choice of the features is still an issue.

The aim of this paper is to develop a robust and fast solution for multiple target detection, labeling and re-identification. It evaluates the behavior of several sets of color and texture/gradient features for covariance matching. In addition, a strategy is proposed for multi-target matching. Note that, contrary to most tracking techniques [7] where the targets have a consistent trajectory, the present work focuses on matching in order to evaluate the descriptors in the context of large motion and object re-identification applications.

The continuation of the paper is structured as follows. Section 2 introduces the covariance matching and the descriptors. Then, Section 3 explains the principles of the objects handling, and how the occlusion, collision and separations are treated in a probabilistic context. To conclude, Section 4 compares the behavior of the features and evaluates the multi-target matching.

2 Covariance Descriptors

2.1 Principles

From each pixel in the observed image I_t of size $W \times H$ a feature vector is obtained by the mapping function ϕ, such that a $W \times H \times d$ dimensional feature image F is generated $F(x, y) = \phi(I, x, y)$, where local information represented by ϕ can be position, color, gradients, filter responses, etc. A rectangular region $\{\mathbf{z}_k\}_{k=1\cdots n}$ of n feature points is represented by the $d \times d$ matrix

$$\mathbf{C}_R = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{z}_k - \mu)(\mathbf{z}_k - \mu)^T \tag{1}$$

where vector μ is the mean of the feature points inside the region. Targets are represented with covariance matrices $\mathbf{C}_R$ which preserve spatial and statistical information and allow to compare different sized regions. Tracking is performed searching for the most similar region in a list of candidate regions in I_t with the object's model in $t-1$. However, direct arithmetic subtraction fails to compare covariance matrices because these type of matrices do not lie on the Euclidean space. The matching can be done applying the dissimilarity measure defined in [8] as the sum of squared logarithms of the generalized eigenvalues. Here, this distance is noted d_{cov}.

Adapting the model for changes of shape, size and appearance is also necessary. This is done by keeping a set of T previous covariance matrices $[\mathbf{C}_1 \cdots \mathbf{C}_T]$ where $\mathbf{C}_1$ denotes the current one, and by computing the mean covariance matrix through Riemannian geometry. A comprehensive explanation of the update mechanism can be found in [9].

2.2 Covariance features

One of our objectives is to test the distinctiveness of several covariance matrices based on descriptors different both in size and nature. Classical features such as

luminance I, image gradients (g_x,g_y), color RGB, and HSV models were tested as well as two color invariants that are worth of interest: the normalized (r, g, b), where r stands for $\frac{R}{R+G+B}$ (similar for G and B) because it offers a separation of luminance and color; then, invariant $L1$ from [10] is interesting because if offers a compact mixture of (r, g, b) and luminance by use of a color relevance measure. Finally, the LBP variance operator VAR_{LBP} is compared to the classical g_x and g_y. Specifically, the tested feature vectors combinations have the following generic form:

$$F_{\mathbf{A},\mathbf{B}} = \left[x\ y\ \mathbf{A}\ \mathbf{B} \right] \tag{2}$$

where $\mathbf{A}$ is a color feature vector and $\mathbf{B}$ a texture descriptor. Five color features $\mathbf{A}$ are tested: Lum (a scalar luminance value), $[R\,G\,B]$, $[H\,S\,V]$, $[r\,g\,b]$, and $[L1\,V]$, as defined in [10] where $L1 = \max(r, g, b)$ and V is the luminance normalized in the object. Two texture descriptors are compared: $grads = [g_x\ g_y]$ is the gradient vector and LBP is the LBP variance value. Thus, ten feature vectors are compared.

2.3 Covariance matching for re-identification

Here, the *Mean Rimemannian Covariance* (**MRC**) matrices proposed by Bak et al. [11] are used to blend appearance information from multiple images.

Given a set of N covariance matrices $\{C_1, C_2, \cdots, C_N - 1\}$, the Karcher or Fréchet mean, is the value μ which minimizes the set of squared distances

$$\mu = arg \min_{C \in \mathcal{M}} \sum_{i=1}^{N} \rho^2(C, C_i) \tag{3}$$

For the case of covariance matrices, the value of μ is calculated iteratively, following the Newton gradient descent method for Riemannian manifolds. The approximate value of μ at step $t + 1$ is

$$\mu_{t+1} = exp_{\mu_t} \left[\frac{1}{N} \sum_{i=1}^{N} log_{\mu_t}(C_i) \right] \tag{4}$$

where, exp_{μ_t} and log_{μ_t} are specific operators uniquely defined on the Riemannian manifold. Equations (5) and (6) express how to calculate them.

$$Y = exp_X(W) = X^{\frac{1}{2}} exp(W) X^{\frac{1}{2}} \tag{5}$$

$$Y = log_X(W) = log(X^{-\frac{1}{2}} W X^{-\frac{1}{2}}) \tag{6}$$

Bak et. al [11] achieve great re-identification rates using a dense grid of of **MRC** matrices. For the case of human signatures, each image is scaled into a fixed size of 64×192 pixels where a grid of overlapping 16×16 pixel size cells is constructed. Neighboring cells are separated by 8 pixel steps. In total, 161 **MRC** are used to construct the human signature.

To reduce the re-identification computational cost we propose a different arrangement of **MRC** matrices. Images are re-scaled to 96×128 pixels, then, rings of concentric rectangles are formed around the image center with exponentially increasing areas allowing some area overlapping. The proposed pattern is inspired in FREAK and DAISY [12, 13] but for rectangular covariance regions, in order to be easily accelerated by the integral images method.

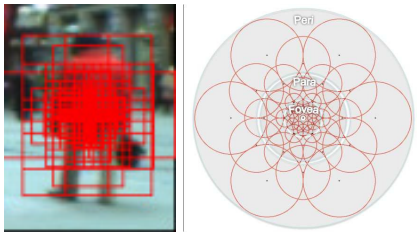

Fig. 1: Proposed **MRC** pattern and its resemblance to FREAK

In our configuration, a total of 42 **MRC** descriptors were employed mostly concentrated at the center while more variability is tolerated at the periphery. To further simplify things **MRC** descriptors are compared one to one in contrast to [11] where comparison is maid sliding one grid against the other.

3 Object Handling

In this section, we describe the multiple-tracking algorithm that we implemented together with the re-identification method (2.3). An approach similar to [14] was followed. A list is kept handling multiple levels of representation: blobs, people and groups. The algorithm receives an image I_t which corresponds to frame t. This image is introduced to the Sigma-Delta [15] background subtraction algorithm where data extracted from the set of previous frames $\{I_0, \cdots, I_{t-1}\}$ is employed to separate foreground and background into a binary image F_t.

Blobs are detected in F_t after applying the Light-Speed Labeling algorithm [16], signaling areas of the image of important change. Blobs of sufficient size are appended to an object list and small blobs are filtered out [1]. The accepted blobs inside the list are now considered objects to match/track exploiting their location, size, trajectory and appearance, modeled with the help of covariance descriptors.

In regular conditions, matching is done considering information of location, size and previous trajectory. Appearance information is used to confirm those estimations, a single low-dimensional (four to six features) covariance descriptor covering the whole object is preferred for simple situations. The complete set of

[1] The parameter used in the paper is 1500px for an image of 768×576

descriptors described in section 4.1 is computed anyway as a preventive measure against faults such as object occlusions, target crossings and objects getting in and out from the scene.

When two or more existing objects become too spatially close, they are merged together to become a *group*. Groups are inserted into a separate list but in general, their location, trajectory and appearance are treated in the same way as any single object. Groups in contrast to single objects, are able to split into separate combinations of their composing original objects.

To each blob in F_t an identity is attributed, which comes from existing or newer objects/groups. As depicted in Fig.3, the identities can transit in five different states: *detected*, *tracked*, *occluded*, *collision* and *lost*. Consider a target blob B, and a set of N candidate objects $\{O_i\}_{i=0\cdots N-1}$, defined by their bounding boxes. For tracking purposes, the euclidean distance d_{bb} between the centers of B and O_i, denoted $\{d_{bb}(B, O_i)\}$, provides a first matching hint. Objects located far from B are filtered out by

$$d_{bb}(B, O_i) > K\max(W, H) \tag{7}$$

where K is an adjustable factor and W, H correspond to the blob's width and height. Note that, when no assumptions can be made on the object location, for multiple-camera object re-identification for example, the location information of (7) is not taken into account.

Consider now a set $Z = \{O_j\}_{j=0\cdots M-1}$, formed by the objects which satisfy inequality (7) where $M \leq N$. A uniform probability $P(O_j) = 1/M$ is assigned to each object. These probabilities are updated considering the evidence provided by the set of distances $D_{bb} = \{d_{bb}(B, O_j)\}_{j=0\cdots M-1}$ as

$$P(O_j|D_{bb}) = \frac{d_{bb}^{-1}(B, O_j)}{\sum_j d_{bb}^{-1}(B, O_j)} \tag{8}$$

Similarly, the set of covariance descriptor distances $D_{cov} = \{d_{cov}(B, O_j)\}_{j=0\cdots M-1}$ allows a second object probability update

$$P(O_j|D_{bb}D_{cov}) = \frac{exp\left(-d_{cov}(B, O_j)\right) P(O_j|D_{bb})}{\sum_j exp\left(-d_{cov}(B, O_j)\right) P(O_j|D_{bb})} \tag{9}$$

the object O_j with the highest posterior probability is assigned to B if it surpasses a minimum threshold.

Groups are formed when multiple objects are merged into one blob, when their spatial distance $d_{bb}(O_i, O_j)$ is low (under a value which depends on the sizes of the two objects).

The covariance descriptors of each individual object are stored before grouping, and additional covariance descriptors are computed for each group, and matched as any other object. When the objects in a group cannot be matched individually with separated new candidate objects, then the matching of the whole group is performed. The individual objects are identified as group members and all of them are set in the state of *collision* sharing the same image location.

Figure 2, displays an example of a merge: at $t = 128$ a descriptor is calculated for the candidate fusion area (red dotted line). Next frame, (due to the closeness) only one blob is detected, and its covariance descriptor matches with the fusion area of previous frame. So, a group is created, it is tracked from frames $t = 130$ to $t = 135$, after this, each object is re-identified individually as described in section 4.1.

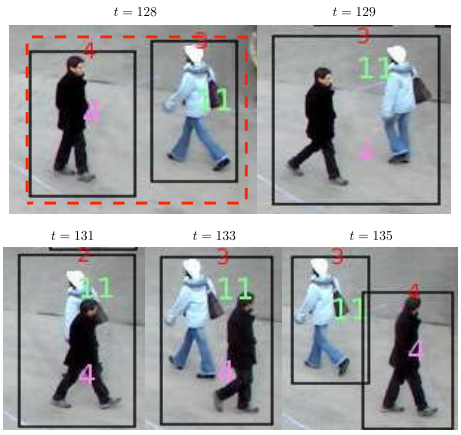

Fig. 2: Example of fusion and separation.

Unmatched blobs are then compared with the objects considered in *collision* or *occluded* at $t - 1$.

Finally, covariance descriptors are regularly updated calculating their covariance mean (equation 3). To avoid model contamination, objects inside a group must not be updated (they are not reliable due to partial occlusion) until they are re-identified individually outside the group. The whole *object handling algorithm* is summarized in **Algorithm 1**.

4 Experiments and results

Our experiments evaluated two different aspects: the re-identification success rate of the proposed **MRC** set for the different feature configurations (details in subsection 2.2), and the proposed multiple-target tracking algorithm.

4.1 Object re-identification experiments

To validate our method of re-identification we used the same performance measure of [11] and [17], which is the Cumulative Matching Characteristic (**CMC**) which represents the percentage of times the correct identity match is found in

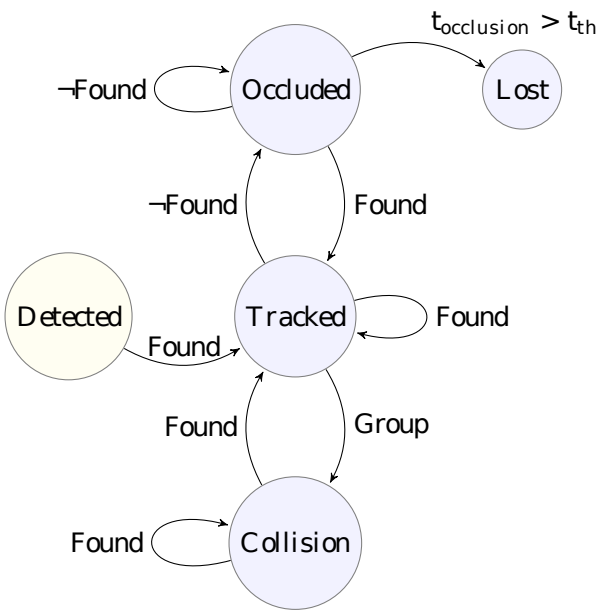

Fig. 3: Objects state transition

the top n matches. Tests were performed for the **ETHZ** [18] and **PETS'09 L1-Walking-Sequence 1** [19] datasets.

ETHZ dataset is composed by images from three different sequences (each one formed by 83, 35 and 28 individuals). Each individual, is captured by the same camera which suits just fine to our single-camera tracking objectives. For this experiment, eight different individuals from **PETS'09 L1-Walking-Sequence 1** were extracted taking discontinuous samples.

For each individual, 10 images were selected from the beginning ant the end and their **MRC** matrices (subsection 4.1) were calculated. The recognition rate was tested, taking random images and comparing against the registered signatures. Care was taken to avoid reusing any of the images occupied during signature calculation. Success is declared when the corresponding image identity is found inside the *top* n list.

To find out which is more discriminant, measurements were taken for the following feature configurations: 1)$F_{lum,LBP}$, 2)$F_{lum,grads}$, 3)$F_{RGB,grads}$, 4)$F_{L1,grads}$ and 5)$F_{HSV,grads}$. Figure 4 reports the results obtained for each sequence.

Except for the first sequence, $F_{RGB,grads}$ is the more powerful configuration to use, achieving good recognition percentages even for the rank-1 score. $F_{L1,grads}$ and $F_{HSV,grads}$ behave similar, because they explicitly separate luminance and colour and they show some resistance to low saturation conditions. On the other side, $F_{lum,grads}$ shows poor recognition performance being overtaken three times by $F_{lum,LBP}$ which has only 4 components.

Algorithm 1: Object handling algorithm

Input: Blobs list *blobs* and Object list *objList*

1 Get blobs covariance descriptors
2 Match blobs - tracked objects
3 Match blobs - candidate collisions
4 Match blobs - occluded and collision objects
5 Non-matched blobs create new objects inside *objList*
6 Dissolve collisions with only one child
7 Remove lost objects
8 Update object states and models
9 Detect candidate collisions

The obtained re-identification rates are comparable to the ones reported in [11] employing a 75% less covariance matrices and fewer components inside them.

4.2 Multiple object tracking

Our tracking algorithm was tested on a randomly walking sparse crowd sequence from the **PETS'09 L1-Walking-Sequence 1 dataset** [19].

Feature vectors proposed in subsection 2.2 are evaluated considering Tracker's Purity (**TP**) [20], which is the ratio of frames a tracker ϵ_i correctly identifies a target $n_{i,j}$ to the total number of frames the tracker exists n_i: $\mathbf{TP} = \frac{n_{i,j}}{n_i}$.

Feature vector combinations which lead to the finest **TP** results were: 1) $F_{RGB,grads}$, 2) $F_{L1,LBP}$, 3) $F_{lum,grads}$, 4) $F_{L1,grads}$, 5) $F_{RGB,LBP}$ and 6) $F_{lum,LBP}$. **TP**s for these combinations are displayed in Figure 6. The points on the circle of radius $\mathbf{TP} = 1$ are related to objects which are always correctly identified in the sequence. The more area is covered, the more often the targets are correctly identified. Mean **TP** values for these combinations are shown in Table 1. Obviously, tracker purity **TP** increases when using color since it provides relevant information that improves distinctiveness. Note that, although $F_{L1,LBP}$ is less distinctive than $F_{RGB,grads}$, it has two advantages: the covariance matrix is more compact (5×5 instead of 7×7) therefore the matching is more rapid, and it offers a better invariance against illumination variations as shown in [10]. The features vectors based on HSV and (r, g, b) are not convincing, since for low saturation the hue is ill-defined and (r, g, b) is not distinctive enough. $F_{lum,LBP}$ degrades severely in comparison to $F_{lum,grads}$, while in the case of the mixture $L1$, the use of VAR_{LBP} does not alter the performances.

5 Conclusions

We have proposed mainly three things in this article: 1) a reduced set of **MRC** matrices which achieves similar to state of the art performances, 2) the incorporation of the VAR_{LBP} operator which produces smaller matrices and 3) a

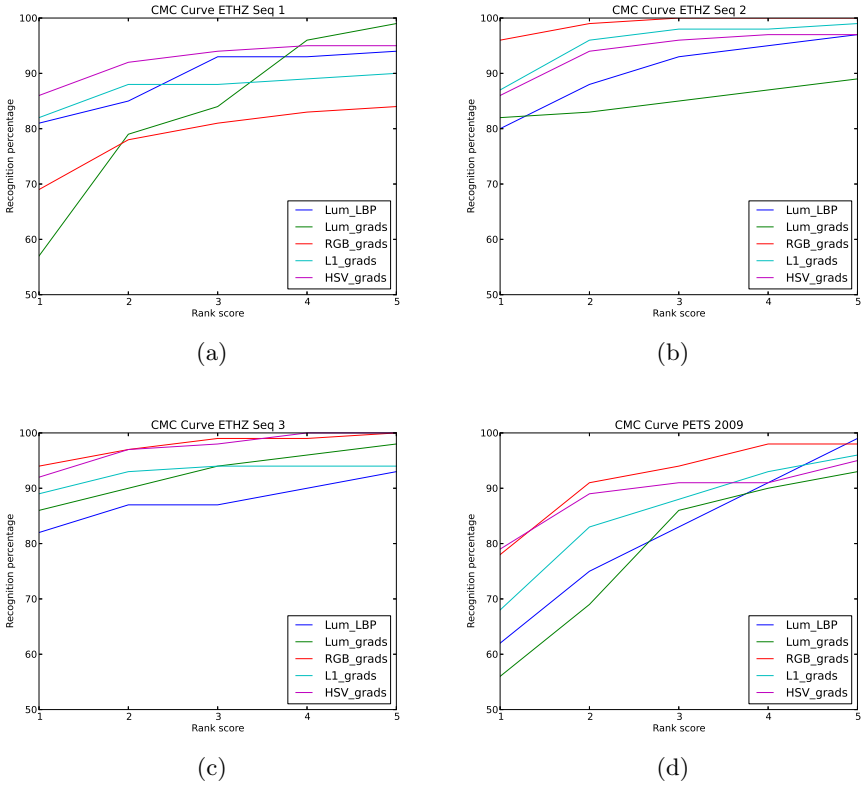

Fig. 4: Cumulative Matching Characteristic **CMC** curves. a) to c) for **ETHZ** sequences and d) for **PETS'09 L1-Walking-Sequence 1**

tracking algorithm which blends localization, trajectory and appearance information providing it with re-identification capabilities.

Here, we have evaluated several descriptors. Note that the use of the invariant $L1$ [10], especially $F_{L1,LBP}$ allows to maintain a performance similar to RGB while being more compact. The use of $L1$ to match images from different cameras and suffering drastic changes of color illumination, will be subject of further investigation.

The proposed multiple-target matching has shown encouraging results. Indeed, although there is intentionally no constraint on the temporal consistency of the trajectories, most objects are correctly matched due to the good distinctiveness of the chosen covariance features. Single crossings between two targets are handled fine regardless of the chosen feature vector combination. Still there are some targets non-consistently identified throughout the sequence (those with low **TP**). This is due to some issues not considered by the model. For example,

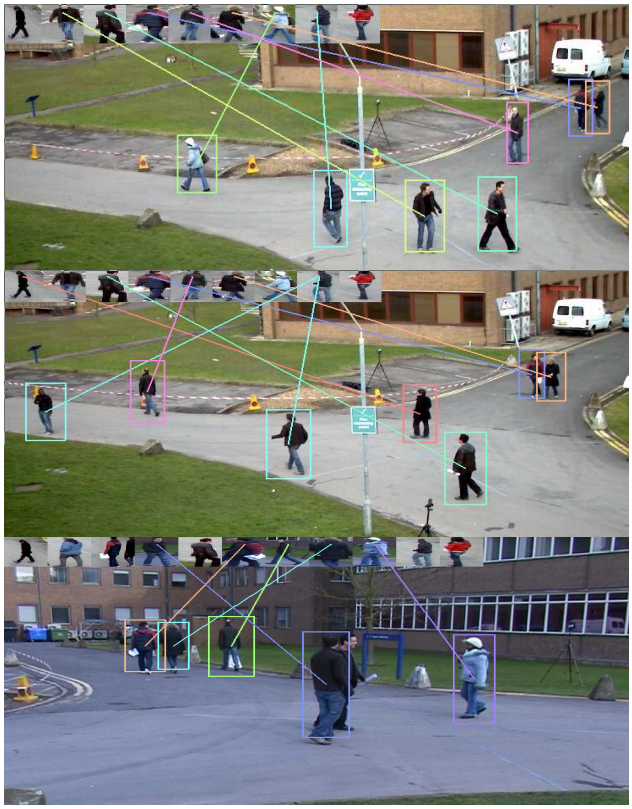

Fig. 5: Some frames of PETS'09 showing the re-identification at different times and points of view.

some problems occur when several targets are crossing each other while some of them experiment background occlusion.

Acknowledgments

This research is supported by the European Project ITEA2 SPY[2].

References

1. Tomasi, C., Kanade, T.: Detection and tracking of point features. Order A Journal On The Theory Of Ordered Sets And Its Applications **7597** (1991) 22
2. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60** (2004) 91–110

[2] Surveillance imProved sYstem http://www.ppsl.asso.fr/spy.php

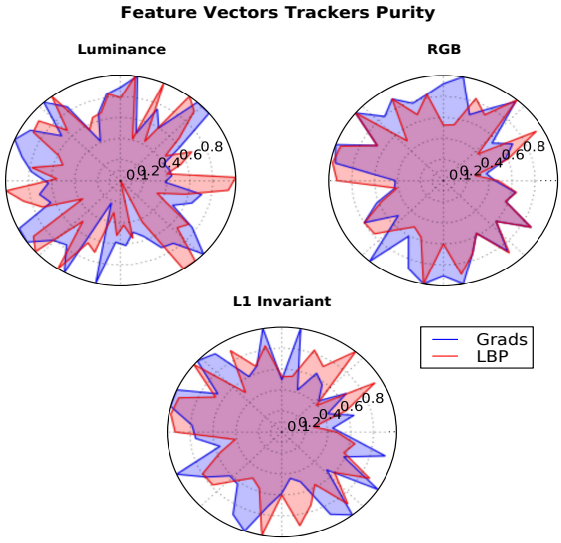

Fig. 6: Trackers Purity **TP** radar plots for different feature vectors.

Space	*Grads*		*LBP*	
	Mean **TP**	FPS	Mean **TP**	FPS
Lum	0.72455	22.887	0.68523	23.628
RGB	0.75046	14.574	0.70516	18.284
L1	0.72209	23.082	0.72823	23.290

Table 1: Mean **TP** and Frames per Second (FPS) associated with the proposed Feature Vectors

3. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell **24** (2002) 603–619
4. Jeyakar, J., Babu, R.V., Ramakrishnan, K.R.: Robust object tracking using local kernels and background information. In: ICIP. (2007) V: 49–52
5. Laguzet, F., Gouiffès, M., Lacassagne, L., Etiemble, D.: Automatic color space switching for robust tracking. In: ICSIPA, IEEE (2011) 295–300
6. Birchfield, S.T., Rangarajan, S.: Spatiograms versus histograms for region-based tracking. In: CVPR. (2005) II: 1158–1163
7. Porikli, F., Tuzel, O., Meer, P.: Covariance Tracking using Model Update Based on Lie Algebra. IEEE CVPR, vol. 1, pp. 728–735, 2006. **1** (2006) 728–735
8. Förstner, W., Moonen, B.: A metric for covariance matrices. Qua vadis geodesia (1999) 113–128
9. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: ICPR. (1994) A:582–585
10. Romero, A., Gouiffès, M., Lacassagne, L.: Feature points tracking adaptive to saturation. In: ICSIPA, IEEE (2011) 277–282

11. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid. In: Advanced Video and Signal-Based Surveillance, Klagenfurt, Autriche (2011)
12. Alahi, A., Ortiz, R., Vandergheynst, P.: FREAK: Fast Retina Keypoint. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012)
13. Tola, E., Lepetit, V., Fua, P.: DAISY: An efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell **32** (2010) 815–830
14. McKenna, S.J., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking groups of people. Computer Vision and Image Understanding **80** (2000) 42 – 56
15. Lacassagne, L., Manzanera, A., Dupret, A.: Motion detection: Fast and robust algorithms for embedded systems. In: Image Processing (ICIP), 2009 16th IEEE International Conference on. (2009) 3265 –3268
16. Lacassagne, L., Zavidovique, B.: Light speed labeling for risc architectures. In: Image Processing (ICIP), 2009 16th IEEE International Conference on. (2009) 3245 –3248
17. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). (2007)
18. Schwartz, W.R., Davis, L.S.: Learning Discriminative Appearance-Based Models Using Partial Least Squares. In: Brazilian Symposium on Computer Graphics and Image Processing. (2009)
19. Ferryman, J., Shahrokni, A.: Pets2009: Dataset and challenge. In: Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on. (2009) 1 –6
20. Smith, K., Gatica-perez, D., marc Odobez, J., Ba, S.: Evaluating multi-object tracking. In: In Workshop on Empirical Evaluation Methods in Computer Vision. (2005)