MULTI-PART TARGET REPRESENTATION FOR COLOR TRACKING

Emilio Maggio and Andrea Cavallaro

Multimedia and Vision Laboratory
Queen Mary, University of London – Mile End Road, London, E1 4NS (United Kingdom)
Email: {emilio.maggio,andrea.cavallaro}@elec.qmul.ac.uk

ABSTRACT

This paper presents an effective target representation based on multiple colour histograms computed on semi-overlapping image areas. This solution introduces spatial information in the representation, without compromising the benefits of the histograms. In particular, target rotation and scaling can be accounted for, thus improving the tracker robustness to false targets. We demonstrate that the proposed target representation outperforms the standard single histogram model and non-overlapping multi-part representations, using state-of-the-art tracking algorithms. Experimental results show that the proposed representation achieves an improvement of tracking accuracy and a reduction of track losses, without increasing significantly the computational complexity.

1. INTRODUCTION

Colour histograms have been widely used to represent, analyse, and characterize images. They allow for significant data reduction, and can be computed efficiently; moreover colour histograms are robust to noise and local image transformations. In the target tracking domain, colour histograms are a popular form of target representation, because of their independence from scaling and rotation, and robustness to partial occlusions [1, 2]. Nevertheless the robustness of such a model is weakened in challenging tasks due to the lack of spatial information. An example is when the searched target state space includes rotation and anisotropic scaling. This problem can be limited by computing more than one histogram on different parts of the target [3], but there is no generally accepted solution for a generic division. A multi-part representation is used in [4] to track ice-hockey players, dividing the rectangular box which bounds the target into two non-overlapping areas, generally corresponding to the shirt and trousers of each player. This solution is effective for the specific application, but is not valid for a generic target.

Since the colour histogram representation relies completely on the information extracted from the target colour distribution, another issue is the choice of an appropriate colour space. Most trackers use a 3D histogram quantizing the RGB colour space [1, 2]. An alternative has been proposed in [3, 4], where HSV is used instead, decoupling chromatic information from shading effects. This model has been employed in [3] to track faces, and in [4] to detect and track ice-hockey players. A large amount of work to provide a reliable color model has been done for skin detection, and Terrillon et al. [5] reviewed skin chrominance models evaluating their performance.

In this paper we propose and evaluate an effective approach based on a multi-part model. The target is divided into overlapping regions in order to increase the tracker sensitivity to rotations and anisotropic scale changes. The proposed partition is independent from the target class, and does not weaken the robustness of the color histogram representation. Moreover, eight color spaces are evaluated and compared in the framework of color histogram based tracker, using an objective quality measure. Three state-of-the-art trackers are used in the tests: the Mean Shift (MS) [1], the Particle Filter (PF) [2], and the Hybrid Tracker (HT) [6]. The first two trackers are widely used in the research community, the third has been recently proposed in [6].

The paper is organized as follows. Section 2 evaluates different color features for target tracking. The multi-part target representation is explained in Section 3. Experimental results are presented in Section 4. Finally, in Section 5 we draw the conclusions.

2. COLOR FEATURES

In order to select the most appropriate color feature, eight color histogram representations are described and evaluated. The eight representations are derived from seven color spaces; RGB, rgb, rg, CIELab, XYZ, YcbCr, and two representations based on HSV (HSV-D, HSV-UC)[1]. These eight different representations are tested on the dataset described in Section 4. The goal is to establish whether there is one performing better, or if it is possible to choose a color space depending on the target class. RGB, XYZ, Lab, and YCbCr are tested. Furthermore two different HSV based representations are implemented. The first is a 3-dimensional HSV histogram created discarding the pixels with $V < 0.1$ and $S < 0.2$ (HDV-D), because under these thresholds the information contained in H is not stable. The second (HSV-UC) is the representation proposed in [3]; HSV-UC decouples the chromatic information from shading, this result is obtained by populating an HS histogram with $N_h N_s$ bins using only the pixels with saturation and value larger than the same two thresholds of HSV-D. The remaining pixels are used to populate N_v additional value-only bins. The total number of bins is $N_h N_s + N_v$, and clearly just N_v pixels contain information related to lighting conditions. The last two color spaces are the normalized *rgb*, and *rg*. In all the cases 10 bins are allocated for each dimension, hence the total number of bins is 1000 for RGB, HSV-D, XYZ, Lab, YCbCr, *rgb*, 110 for HSV-UC, and 100 for *rg*. The MS algorithm described in [1] is employed in the tests. MS is chosen because is parameter free.

The performance evaluation is based on a metric using true positive pixels $TP(i)$ in each frame i. The number of true positives is the number of pixels belonging both to the ground truth

[1] http://www.poynton.com/PDFs/coloureq.pdf .

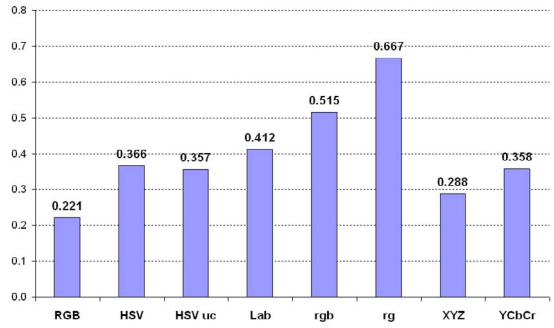

Fig. 1. Comparison of tracking accuracy results (average distance from the ground truth), using eight different color representations.

ellipse, as well as to the tracker output. The metric is defined as

$$OD(i) = 1 - \frac{2 \times TP(i)}{Card(\, A_c(i)\,) + Card(\, A_{gt}(i)\,)}. \qquad (1)$$

where $A_{gt}(i)$ and $A_c(i)$ are the ground truth and the candidate area, respectively. This normalized metric rewards candidates with a high percentage of true positive pixels, and with few false positives and false negatives. Using Eq.(1) a Lost Track (LT) is declared at the frame i when $OD(i) > 0.8$.

Tab. 1 presents the summary of the results. It is possible to notice that some color spaces have proved to be inadequate to a tracking purpose: *rg* and *rgb*. By discarding the luminance information, the tracker results in several lost tracks and in a poor quality of the trajectory. The results achieved by XYZ are quite close to RGB (in one case better), due to the similarity between the information carried on. The two representations based on HSV achieve good results in particular conditions: HSV-D outperforms RGB in three sequences but is completely unreliable when objects with low saturated colors are tracked; HSV-UC outperforms RGB two times in sequences with variable lighting conditions and target self-shadowing. This representation proves to be efficient, especially considering that the number of bins is reduced of an order of magnitude.

The average results calculated over the dataset (Fig. 1) show that the RGB-based representation outperforms the others. RGB is the only color space which does not results in lost tracks, and achieves the best average score on the entire dataset. Hence, for a general application with different target classes, we choose RGB.

3. MULTI-PART TARGET REPRESENTATION

The lack of spatial information in histograms can be a problem in color based tracking as showed in the top row of Fig. 4. The tracker (PF) is attracted to false targets with similar color distributions. For instance in Fig. 4 (a)-(c), first the car shadow is black as the trousers, and then the car is white as the shirt. In this case, the information related to the spatial distribution of the colors is fundamental for a correct tracking. Moreover, when more precise estimation of the target orientation and size is necessary, spatial information is beneficial as well. In this section we propose a solution to introduce this information in the target representation.

In the case of ellipse-based trackers the proposed representation is showed in Fig. 2. Seven histograms are calculated over

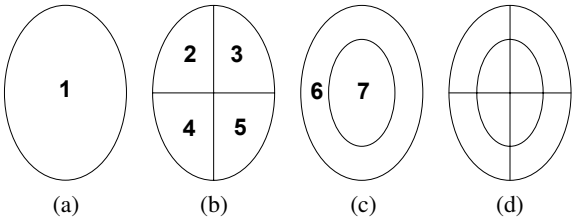

Fig. 2. Multi-part representation. (a) Whole, (b) rotation sensitive division, (c) size sensitive division, (d) target overall division.

semi-overlapping regions of the ellipse. The first histogram is calculated on the whole ellipse. Four parts are obtained from the ellipse partition created by the two axes to add spatial information necessary to recognize rotations of the target. Other two parts are the inside and outside area of an ellipse with same eccentricity, but half axis size then the whole ellipse. The proposed model based on this partition will be referred in the following as 7MP.

To calculate the likelihood of a candidate a function which defines a distance between the model and the candidate is needed. A commonly used metric is based on the Bhattacharryyan coefficient [7]. The distance between two m-bins normalized histograms a and b is defined as

$$d\,[a,b] = \sqrt{1 - \sum_{u=1}^{m} \sqrt{a_u, b_u}}. \qquad (2)$$

Hence the distance between the multi-part model q and the candidate $p(\mathbf{y})$ with ellipse parameters $\mathbf{y}$, is calculated using the average of the Bhattacharryyan distance as

$$\rho\,[p(\mathbf{y}), q] = \frac{\sum_{i=1}^{N} d\,[p^i(\mathbf{y}), q^i]}{N}, \qquad (3)$$

where p^i and q^i are the model and candidate histograms calculated on the $i-th$ sub-part of the ellipse, and N is the number of sub-parts (in our case $N = 7$).

The multi-part representation 7MP is compared with other partitions of the target in Tab. 2; the experiments used HT with the 3-dimensional state space. The representations under analysis are the following: the two-part horizontal division (2HD) proposed in [4], then the two-part division (IN-OUT) of Fig. 2 (c), and the four parts (4PARTS) of Fig. 2 (b). These partitions divide the target into non overlapping areas and do not include the histogram calculated on the whole ellipse. 7MP outperforms the other multi-part representations due to the more complete spatial information included. For example comparing 7MP with 2HD, 7MP introduces information discriminant to object rotations. One of the 7MP advantages is that it is not designed for a particular target class. This representation maintains the flexibility and robustness to occlusions of the color histogram, and improves the performance of a the single-part based tracker. Moreover the six extra histograms can be calculated efficiently, adding just the operations necessary for the histogram selection.

4. RESULTS

The results presented in this section are based on a dataset of targets extracted from 5 different test sequences (Fig. 3). They can be

Table 1. Comparison of tracking accuracy results for eight different colour representations on three target classes.

		RGB	HSV-D	HSV-UC	Lab	rgb	rg	XYZ	YCbCr
PEOPLE	P1	0.122	0.218	0.173	LT	LT	LT	0.131	0.162
	P2	0.437	0.407	0.393	0.448	LT	LT	0.429	0.638
FACES	F1	0.116	0.097	LT	0.213	LT	LT	0.156	0.268
	F2	0.090	0.102	0.121	0.167	0.140	0.179	0.115	0.175
	F3	0.332	0.309	0.253	LT	0.301	LT	LT	LT
VEHICLES	V1	0.302	LT	LT	LT	LT	LT	0.461	LT
	V2	0.121	0.140	0.146	0.119	0.261	LT	0.128	0.117
	V3	0.319	LT	0.409	LT	LT	LT	0.322	0.309
	V4	0.154	0.229	0.208	0.212	0.302	LT	0.204	0.161

Table 2. Comparison of tracking accuracy results for different multi-part divisions.

	2HD	IN-OUT	4PARTS	7MP
P1	0.111	0.114	0.115	0.108
P2	0.204	0.223	0.164	0.145
F1	0.133	0.134	0.133	0.131
F2	0.314(LT)	0.288	0.216	0.162
F3	0.304	0.523(LT)	0.230	0.203
V4	0.316	0.246	0.325	0.289
Average	**0.231**	**0.255**	**0.197**	**0.172**

divided into three classes: (i) PEOPLE: two pedestrians from the PETS2001 DATA-SET1 (P1: man with backpack, P2: man with gray pull), and a person walking (P3) from the sequence *Walk2* (EC Funded CAVIAR project IST 2001 37540). (ii) FACES: two from high quality sequences *Toni* (F1) and *Nikola* (F2), and one from low quality sequence *Emilio* (F3). (iii) VEHICLES: three cars (V1: small blue car, V2: blue car, V3: white van), and a truck (V4) are selected from the CIF sequence *Highway*, and a car (V5) from the PETS 2001 DATA-SET1.

The parameters are the same for all test sequences, and are described in the following. The histograms are calculated in the RGB space with 10x10x10 bins. MS runs 5 times with different kernel sizes up to +/-10% than the previous frame, and uses the 3-dimensional state model. PF and HT use both 5-dimensional (5D) and 3-dimensional (3D) state models. The state model is composed of the target position, (x, y), and the target size h in the 3D case, while ellipse eccentricity e, and rotation θ are added in the 5D case. PF uses a zero-order motion model with fixed $\sigma_x = \sigma_y = 6.5$, $\sigma_h = 0.05$, $\sigma_e = 0.02$, and $\sigma_\theta = 3.5^o$. (The scale change is a percent, the position is in pixels, and the angle in grades). The initial values for the adaptive state transition model employed with HT are: $\sigma_x^0 = \sigma_y^0 = 12$, $\sigma_h^0 = 0.1$, $\sigma_e^0 = 0.04$, and $\sigma_\theta^0 = 10^o$. PF uses 150 samples and HT 30. 7MP applied to different algorithms is compared using the metric defined in Eq. (1).

Fig. 4 (a)-(f) shows the tracking results of PF with 7MP (Fig. 4 (d)-(f)), and without (Fig. 4 (a)-(c)). The tracker using 7MP is not attracted by false targets with similar color properties: the spatial information introduced in the model (i.e., the relative position of the shirt and trousers) avoids the lost track improving the overall average quality (Tab. 3). The same considerations can be done for the pedestrian P2 (Fig. 4 (k)-(n)). This target proposes challenging conditions for a tracker, due to the small size and to several false targets nearby the trajectory. The results of the tracking with MS are showed in Fig. 4 (k)-(n). 7MP avoids the lost track (Fig. 4 (m)(n)) produced after few frames by the standard representation (Fig. 4 (k)(l)).

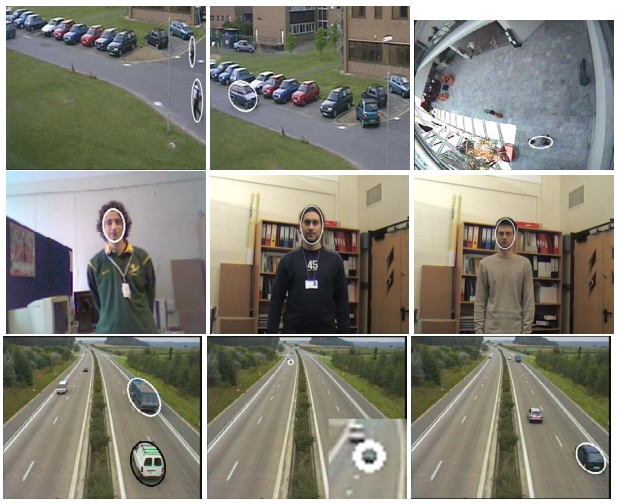

Fig. 3. Target initialization. Top row: three pedestrians (P1,P2,P3) and a car (V5) from PETS databases. Middle row: faces on cluttered background (F1,F2,F3). Bottom: three cars and a truck from the MPEG-7 test sequence *Highway* (V1,V2,V3,V4).

In Fig. 4 (g)-(j) sample frames of F1 face tracking using HT and 5D state space are showed. The single-part representation produces a wrong result in terms of target orientation and size (Fig. 4 (g)(h)); the face and the bookshelf in the background have similar colors, and the representation is not able to distinguish correctly the face. On the contrary, spatial information contained in the proposed model solves this problem (Fig. 4 (i)(j)); in this case the information related to the relative position between the hairs and the skin makes the difference.

The complete results are summarized in Tab. 3. By analyzing these results, it is possible to notice that the multi-part representation improves in average the performances of all the three algorithms (MS, PF and HT). The only target where the standard representation outperforms 7MP is the blue truck of the sequence *Highway* (V4). This result is due to the fact that the target becomes smaller and smaller, hence it is useless to split the few pixels remaining into different histograms. Moreover, V4 does not present statistically different color properties, and in this case 7MP is not effective. The size problem could be solved by using a size threshold under which the single histogram representation is used. Moreover, it would be possible to analyze the target color layout at the initialization to decide whether to use the multi-part representation. The algorithm which benefits more of 7MP is PF; it gains in

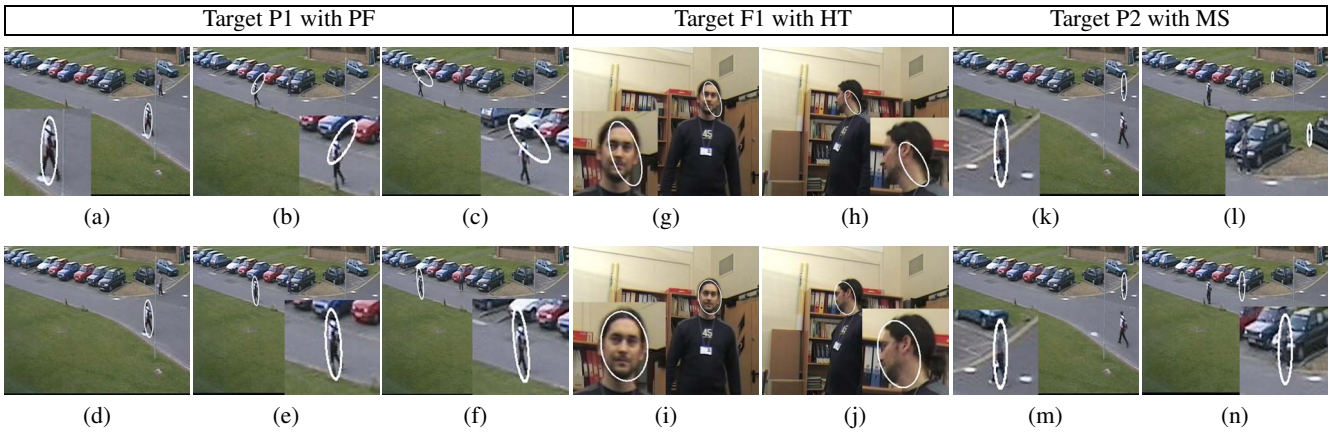

Fig. 4. Examples of tracking results using single-part representation (top row), and multi-part representation (bottom row). Target P1 (frames 1500,1710,1721) tracked with Particle Filter; target F1 (frames 50,115) tracked with Hybrid Tracker; target P2 (frames 1461,1565) tracked with Mean Shift. The videos with the results are available at http://www.elec.qmul.ac.uk/staffinfo/andrea/MP.html

Table 3. Comparison of tracking accuracy results for different target representations.

	5D state space					**3D state space**					
	SP		**7MP**			**SP**			**7MP**		
	PF	**HT**	**PF**	**HT**		**MS**	**PF**	**HT**	**MS**	**PF**	**HT**
P1	0.346(LT)	0.364(LT)	0.109	0.125		0.125	0.138	0.119	0.145	0.105	0.108
P2	0.838(LT)	0.657(LT)	0.227(LT)	0.644(LT)		0.84(LT)	0.351(LT)	0.348(LT)	0.163	0.198(LT)	0.145
F1	0.118	0.101	0.101	0.115		0.161	0.104	0.158	0.129	0.100	0.131
F2	0.304(LT)	0.297(LT)	0.274(LT)	0.284(LT)		0.188	0.295(LT)	0.192	0.158	0.270(LT)	0.162
F3	0.460(LT)	0.393	0.228	0.262		0.567(LT)	0.440(LT)	0.357	0.456(LT)	0.197	0.203
V4	0.212	0.200	0.258	0.282		0.160	0.231	0.176	0.262	0.305	0.289
V5	0.257	0.260	0.165	0.139		0.449(LT)	0.336	0.410(LT)	0.386(LT)	0.189	0.393(LT)
Average	**0.362**	**0.324**	**0.194**	**0.265**		**0.356**	**0.271**	**0.251**	**0.243**	**0.195**	**0.203**

average about 38% in terms of quality reducing the number of lost track from 7 to 4.

5. CONCLUSIONS

We presented a multi-part target representation based on the computation of seven semi-overlapping color histograms. After evaluating eight different color spaces, the RGB color space was selected and applied to the multi-part representation; which has been tested with three different algorithms, namely Particle Filter, Mean Shift, and Hybrid Tracker. Experimental results showed that the proposed representation is more accurate than the single histogram and multiple non-overlapping ones, and achieves better results in predicting the correct orientation and size of the target.

Future work includes investigating an appropriate adaptive representation based on target content, in order to weight differently the parts of the representation.

6. REFERENCES

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.

[2] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "A color-based particle filter," in *Proc.of the 1st Workshop on Generative-Model-Based Vision*, June 2002, pp. 53–60.

[3] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. of the European Conference on Computer Vision*, Copenhagen, May-June 2002, vol. 1, pp. 661–675.

[4] K. Okuma, A. Taleghani, N. De Freitas, J.J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proc. of the European Conference on Computer Vision*, Prague, May 2004.

[5] J.C. Terrillon, M.N. Shirazi, H. Fukamachi, and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," in *Proc. of IEEE International Conf. on Automatic Face and Gesture Recognition*, Grenoble, Mar. 2000, pp. 54–61.

[6] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. of IEEE International Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, 2005.

[7] T.Kailath, "The divergence and bhattacharryya distance measures in signal selection," *IEEE Trans. Comm. Technology*, vol. 15, pp. 52–60, 1967.