Dynamic Multi-cue Tracking with Detection Responses Association

Guochen Jia[1], Yonghong Tian[1], Yaowei Wang[2], Tiejun Huang[1], Min Wang[1]
[1]National Engineering Laboratory for Video Technology, School of EE & CS, Peking University, Beijing 100871, China
[2]Beijing Institute of Technology, Beijing 100081, China
{gcjia, ywwang, mwang}@jdl.ac.cn, {yhtian, tjhuang}@pku.edu.cn

ABSTRACT
Multi-cue integration has proved successful at increasing the robustness of tracking algorithms and overcoming the failure cases of individual cue. But considering dynamic appearance of objects or clutter background, the integration based on constant weights may weaken the performance of this scheme. In this paper, we propose a dynamic weights update mechanism for multiple cues tracking with detection responses as supervision. We integrate multiple cues based on the observation hypotheses compared with detection association results and adjust the weights according to the approximation degree. The integration is adapted on-the-fly during tracking, in order to keep the tracker adaptive. The proposed method allows flexible combination of different cues and we select cues based on color and local feature for tracking. Experiments are carried out on 602 trajectories extracted from TRECVID 2008 event detection dataset which is recorded in an airport scenario. Comparison results prove the effectiveness of our method.

Categories and Subject Descriptors
I.4.8 [**IMAGE PROCESSING AND COMPUTER VISION**]: Scene Analysis - *Tracking.*

General Terms
Algorithms, Experimentation.

Keywords
Multiple cues integration, Dynamic appearance, Detection association.

1. INTRODUCTION
Tracking video objects is an important issue in scene analysis, and has lots of variety applications including video surveillance, human-computer interaction, driving assistance and video compression. Even though there are lots of significant progress, object tracking is still persecuted by many challenges like clutter background, dynamic object appearance, significant occlusion and so on, which make tracking results instable and inaccurate.
In order to cope with those difficulties, many algorithms have been proposed [1]. However, most of them only use an individual cue and solo cue can't provide all-around invariance to different scene conditions. So it's appealing to integrate multiple complementary cues into one observation model for tracking. Some works focus on using probabilistic approaches to model interactions among multiple cues [2, 3, 4]. Considering the cues' dependence, Wu *et al*. [2] solved the problem based on graphical model like Hidden Markov Model and proposed a co-inference process to illustrate the interactions among different cues. Similarly, Du *et al.* [3] performed tracking in different cues by HMM. Then Chain's models were used to link these parallel HMMs and to represent the dependence between these cues. Moreno-Noguer *et al*. [4] designed different Bayesian filter for each cue and considered the mutual dependence during hypotheses correction stage. Another approach that has gained popularity recently is to integrate the cues based on the discriminability between foreground object and background region. Collins *et al.* [5] emphasized the importance of the background appearance, and adaptively selected top-ranked discriminative features for tracking. Yang *et al.* [6] defined dissimilarity function for each cue according to its discriminative power and applied regression process to adapt the integration of multiple cues. Babenko *et al*. [7] utilized on-line Multiple Instance Learning approach to update the appearance model with a set of image patches and selected the best features from a fixed pool of features.
Recently, tracking associated with detection has become popular because involving detection responses could help to locate object exactly and alleviate drift [9, 10, 11]. Frame-to-frame detection responses are linked via cues like position, appearance and size in [9]. Breitenstein *et al*. [10] integrated detector into tracking by monitoring continuous detection confidence in particle filtering framework. Through online-trained classifier, prior knowledge is introduced and reliable detections are associated in the end. Xing *et al*. [11] collected detection responses in a temporal sliding window and associated them with potential tracklets. The association is also based on position, size and appearance. Although detection-based tracking methods have achieved lots of progress, the association process based on single feature or fixed features set, could be further exploited.
In this paper, we present a dynamic multi-cue tracking scheme by integrating color and local feature and introduce Histograms of Oriented Gradients (HOG) detection responses to guide the integration. This paper firstly proposes a method to evaluate each cue based on approximation of observation between hypothesis and tracking result. The approximation is measured through condition probability while the observations could be obtained from different stand-alone tracking algorithms. Secondly this paper introduces reliable detection responses as supervision to

MM'10, October 25-29, 2010, Firenze, Italy.

guide the integration. One response could be associated with one tracking hypothesis if the confidence measured by features set is above certain threshold. The work most related to ours is that of [8]. The difference from the previous one is that we evaluate cues while tracking and on-line adjust weights whereas in [8] weights are learned on an off-line training set. Moreover, we associate detection responses based on dynamic features set to enhance tracking performance while similar works often use fixed set. Figure 1 illustrates the block diagram of our scheme.

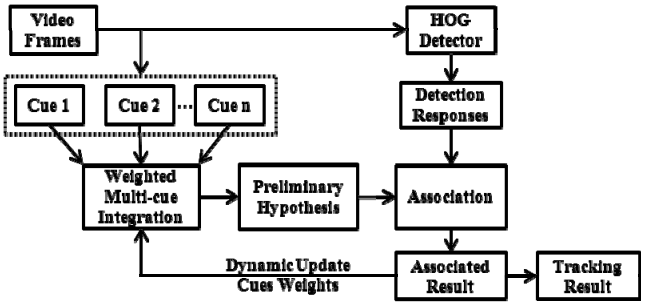

Figure 1. Block diagram of proposed tracking scheme

2. DYNAMIC MULTI-CUE INTEGRATION

When an object moves in the scene, its motion states are represented in image observations. If we denote object states and image observations by X_t and Z_t respectively, the relationship can be expressed as $P(X_t \mid X_{t-1}) \propto P(Z_t \mid Z_{t-1})$. Given several hypotheses in *t th* frame, the tracking task can be formulated as finding the MAP estimation $z_t^* = \arg\max_{z_t} p(z_t \mid z_{t-1})$. We could establish one observation model for one cue and different cues may lead to various observation distributions. Sole feature only works well in specific scene, so there is a need to integrate multiple observations into one robust model.

For each cue, we establish an observation model and get the condition probability distribution $p_i(z_t \mid z_{t-1})$. This value could be obtained through kinds of methods, e.g. a matched point ratio, a histogram distance, or a classification margin. According to the distribution of this cue, the tracking observation hypothesis should be $z_t^i = \arg\max_{z_i} p_i(z_t \mid z_{t-1})$. Given final tracking result z_t^*, we can get equation as follows:

$$p_i(z_t^* \mid z_{t-1}) = p_i(z_t^* \mid z_t^i) p_i(z_t^i \mid z_{t-1}), \tag{1}$$

where $p_i(z_t^* \mid z_t^i)$ represents the measurement of observation approximation. A greater value means a better estimation of z_t^*, and the approximation degree could be used for evaluating cue *i*. Inspired by this, the weight w_i assigned to *i th* cue should reflect this relationship denoted by $w_i \propto p_i(z_t^* \mid z_t^i)$. On the other hand, spatial distance also reflects the approximation degree. A closer hypothesis obviously implicates a better estimation based on the cue. This metric can be calculated through bounding box overlap or a scale normalized distance between two centers. If we represent the spatial distance as $d(z_t^* \mid z_t^i)$, weights should satisfy the property like $w_i \propto 1/d(z_t^* \mid z_t^i)$. We normalize the distance, and hence the relationship is denoted as $w_i \propto e^{-d(z_t^* \mid z_t^i)}$. In conclusion, weights of cues should reflect both the observation approximation and the spatial distance. We collect *n* hypothesis corresponding to *n* cues, and evaluate *i th* cue as follows:

$$w_i = C \times p_i(z_t^* \mid z_t^i) e^{-d(z_t^* \mid z_t^i)}, \tag{2}$$

where $C = 1/\sum_{i=1}^{n} p_i(z_t^* \mid z_t^i) e^{-d(z_t^* \mid z_t^i)}$ is a normalization constant. During real tracking process, each cue generates one hypothesis in parallel, and a linear weighted method is employed for the multi-cue integration. We combine the outputs of all components and predict the tracking result as $z^* = \sum_{i=1}^{n} w_i z^i$. Once the weight of one cue becomes zero, its hypothesis will be discarded.

If the ground truth is available, we can obtain all exact tracking results z^* in advance and evaluate the performance of each cue like works in [8]. In this way, the weights are learned off-line and fixed during real tracking process. As for our multi-cue tracking scheme, we introduce detection association process which will be discussed in Section 4, and use the associated result $\hat{z}_t$ as the estimation of z_t^*. Once the associated result in each frame is available, it can be used to guide the integration for next frame. According to this, we can adapt the integration on-the-fly during tracking and achieve the goal of dynamic update.

3. CUES FOR TRACKING

3.1 Color

Various cues have been used for object tracking. Each of them has its pros and cons but the color has gained the most of attention as it is well distinguishable to human eye and seems to contain a good amount of useful information. In this paper we employ kernel-weighted color histogram to describe the color properties of objects [12], which consists of 16 × 16 × 16 bins in the RGB space. Bhattacharrya coefficient is used to measure the condition probability between observations as follows:

$$\begin{aligned} p_{color}(z_t \mid z_{t-1}) &= H_{z_t}^B H_{z_{t-1}}^B = \rho[\hat{p}(z_t), \hat{q}(z_{t-1})] \\ &= \sum_{u=1}^{B} \sqrt{\hat{p}_u(z_t)\hat{q}_u(z_{t-1})}, \end{aligned} \tag{3}$$

where *B* is the entry number of histogram. Then mean-shift [12] method is applied to find the peak of this distribution and generate hypothesis z_t^{color}.

3.2 Local Feature

Inspired by recent progress in object recognition based on local descriptors, we introduce SURF [13] as a complementary cue for color. Interesting points are used to estimate inter-frame motion as well as scale change. Here, the condition probability of two observations is defined as follows:

$$p_{local}(z_t \mid z_{t-1}) = \frac{2 \times T_{\cap}(z_{t-1}, z_t)}{T(z_{t-1}) + T(z_t)}, \tag{4}$$

where $T(z_{t-1})$ and $T(z_t)$ are numbers of interesting points in two hypothesis regions respectively, and $T_{\cap}(z_{t-1}, z_t)$ is the number of matched points between two regions. During tracking process, we suppose the correlation between consecutive frames is fairly high, and use a fixed-size search window. Pairs of matched points are used to estimate X-axis and Y-axis shifts. Let $\bar{\sigma}_m$ be the average scale of the matched SURF points in observation *m* and $\bar{\sigma}_c$ be the average scale of the matched SURF points in observation *c*. We apply equation (5) to estimate the scale, similar to [14]:

$$s_c = s_m \times (1 + \frac{\bar{\sigma}_c - \bar{\sigma}_m}{(1+\theta) \times \bar{\sigma}_m}), \quad (5)$$

where s_c and s_m are the scales of observation and $\theta \in [0, +\infty)$ is used to control the scale change.

4. DETECTION ASSOCIATION

Before the association, we have already obtained preliminary hypothesis for tracked object, then we need to associate detection responses around it to determine final tracking result. Detection responses are collected using a HOG detector. In this paper, we train a high precision with acceptable recall detector to provide reliable information. All detection responses are denoted as $D^t = \{d_1^t, d_2^t, ..., d_l^t\}$. Given one hypothesis $p^t = (x^p, y^p, s^p)$ and one detection response $d_i^t = (x^d, y^d, s^d) \in D^t$, where (x, y) is the translation and s is the scale, we calculate the association likelihood between them as

$$l_i(d_i^t, p^t) = \sum_{r=1}^{n} w_r^{t-1} p_r(d_i^t \mid p^t), \quad (6)$$

where w_r^{t-1} is the weight of *r th* cue at *t*-1 *th* frame and $p_r(\cdot)$ is the condition probability function described in Section 3. Pairwise likelihoods above certain threshold are collected and the maximum one is selected as the final assignment. As for multiple objects tracking, supposing there are *l* detection responses and *n* preliminary hypotheses, we can get an association likelihood matrix as follows using equation (6):

$$\begin{bmatrix} l_{1,1} & l_{1,2} & l_{1,3} & \dots & l_{1,n} \\ l_{2,1} & l_{2,2} & l_{2,3} & \dots & l_{2,n} \\ \dots & \dots & \dots & \dots & \dots \\ l_{l,1} & l_{l,2} & l_{l,3} & \dots & l_{l,n} \end{bmatrix}_{l \times n} . \quad (7)$$

All pairwise likelihoods above certain threshold are collected and the classical Hungarian algorithm is used to obtain optimal single-frame assignment. If d_i^t cannot be associated with others, it will be used to initialize a new tracker. On the other hand, considering the associated observations z_p and z_d, we can get:

$$\hat{z}_t = \lambda z_p + (1-\lambda) z_d = \lambda (x^p, y^p, s^p) + (1-\lambda)(x^d, y^d, s^d). \quad (8)$$

The factor $\lambda \in [0,1]$ is used to supervise tracking. We tested five difference values (0.0, 0.05, 0.1, 0.15, 0.2), and chose 0.15 for the best results. During the association, this paper utilizes cues from color and local feature. Differing from previous works, the association is based on dynamic features set which is determined during tracking process. In this way, we adaptively associate tracking hypothesis with detection responses.

Once the associated result $\hat{z}_t$ is obtained, we use it as an estimation of z_t^* and adjust the multi-cue integration described in Section 2. Therefore, we realize the dynamic multi-cue tracking scheme.

5. EXPERIMENTAL RESLUTS

5.1 Experiment Setup

In order to evaluate our tracking methods, we carry out experiments on public TRECVID 2008 Event Detection dataset [15]. The dataset is collected in an airport scenario with frame in size of 720 ×576 and frame rate of 25fps. This dataset is very challenging because of the heavy inter-person occlusions and poor image contrast between objects and background. We cut 43 video clips and labeled human head-shoulders every four frames from one view as the ground truth. 602 trajectories whose lengths are between 40 frames and 1000 frames are chosen for the experiments. In the experiments, we train a high precision (around 80%) with acceptable recall (around 30%) HOG detector to provide detection responses. The detector is run individually and detection results for each frame are recorded before tracking.

5.2 Performance Evaluation

In order to evaluate tracking performance, we adopt the metrics in the CLEAR evaluation [16] and select five adopted metrics:

1) GT: Ground Truth. Number of ground truth trajectories.
2) MT: Mostly tracked. Number of GT trajectories which are covered by tracker output for more than 80% in length. The greater the better.
3) ML: Mostly lost. Number of GT trajectories which are covered by tracker output for less than 20 % in length.
4) PT: Partially tracked. Number GT trajectories which are covered by tracker output between 80% and 20%.
5) Frag: Fragments. The total number of fragments of ground truth trajectories in tracking result. The smaller the better.

And these metrics can be traded off between accuracy and consistency of tracking.

For each trajectory, we use the ground truth to initialize tracker and run tracking processes till the end of the trajectory. Once the tracker loses the object, it will be re-initialized. Our method is compared with the method the methods using single cue based on color and SURF information. In addition, we also employ the cascaded strategy described in [8]. In our multi-cue scheme, the initial weights of color and SURF are set at 0.5. Table 1 shows the comparison results. Among the four, our method successfully tracks the most trajectories and achieves best results.

Table 1. Results on TRECVID 08 event detection dataset

Cues	GT	MT	ML	PT	Frag
Color [12]	602	125	232	245	1011
SURF	602	88	230	284	1020
Cascaded [8]	602	121	227	254	1023
Our Method	602	274	105	223	1015

In Figure 2 we illustrate some sample results. Video recorded two men walking toward the gats under illumination variation and clutter background. The red box distracted tracker based on color cues because its appearance is similar to object close-by. Background near the gate is clutter which deviated and failed the tracker based on SURF information. And the cascaded tracker didn't overcome these difficulties, either. As we can see, our method could track persons robustly by taking advantage of the multi-cue integration and the detection association.

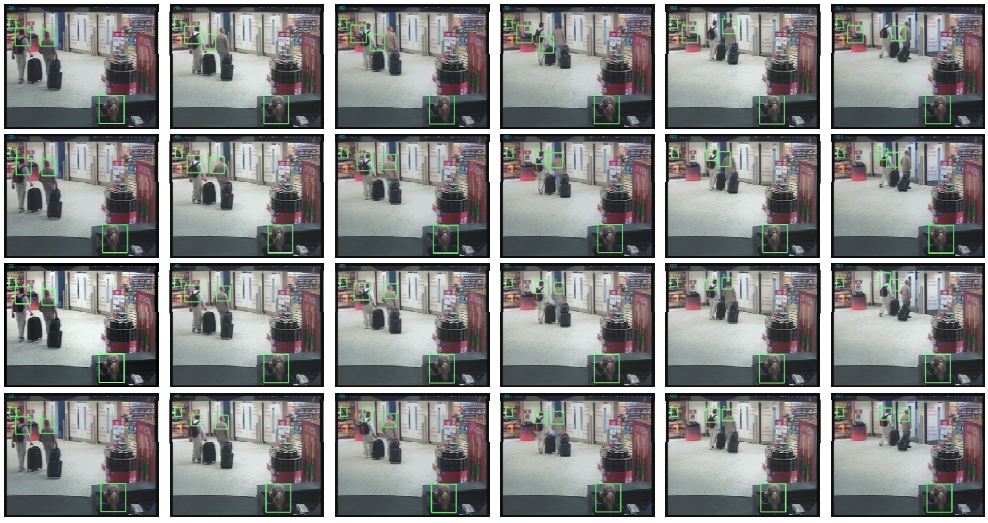

Figure 2. Tracking results on TRECVID08 dataset. First row are results based on color and second row based on local feature. The third row shows results of method employing cascaded strategy, and the bottom row are our method's results.

6. CONCLUSION

In this paper, we propose a dynamic weights update mechanism for multiple cues tracking which introduces HOG detection responses as supervision. Each cue is integrated with respect to its observation compared with detection association result and the integration is adapted on-the-fly during tracking. Color and local feature cues are employed in experiments and comparison results prove the effectiveness of our method. Our future work will include the investigation of the existing relationships between different cues and dependent multi-cue integration for object tracking.

7. ACKNOWLEDGMENTS

This work is supported by grants from the Chinese National Natural Science Foundation under contract No. 60973055 and No. 90820003, National Basic Research Program of China under contract No. 2009CB320906, and Fok Ying Dong Education Foundation under contract No. 122008.

8. REFERENCES

[1] Yilmaz, A., Javed, O., and Shah, M. Object Tracking: A Survey. ACM Comput. Surv. vol.38, no.4, Article 13 (Dec.2006), 45 pages.

[2] Wu, Y., and Huang, T. S. Robust visual tracking by Integrating Multiple Cues based on Co-inference Learning. Int'l Journal of Computer Vision, vol 58(1):55-71, 2004.

[3] Du, W., and Piater, J. A Probabilistic Approach to Integrating Multiple Cues in Visual Tracking. In: Pro of the Euro. Conf. on Computer Vision, 2008.

[4] Francesc Moreno-Noguer, Sanfeliu, A., and Samaras, D. Dependent Multiple Cue Integration for Robust Tracking. IEEE Trans. PAMI, vol 30(4): 670-685, 2008.

[5] Collins, R. T., Liu, Y., and Leordeanu, M. Online Selection of Discriminative Tracking Features. IEEE Trans. PAMI, vol 27(10): 1631-1643, 2005.

[6] Yang, M., Lv, F., Xu, W., and Gong, Y. Detection Driven Adaptive Multi-cue Integration for Multiple Human Tracking. in Proc. Int. Conf. Computer Vision, 2009.

[7] Babenko, B., Yang, M-H., and Belongie, S. Visual Tracking with Online Multiple Instance Learning. in Proc. of Computer Vision and Pattern Recognition, 2009.

[8] Stenger,B., Woodley, T., and Cipolla, R. Learning to Track with Multiple Observers. in Proc. Int. Conf. Computer Vision and Pattern Recognition, 2009.

[9] Huang, C., Wu, B., and Nevatia, R. Robust Object Tracking by Hierarchical Association of Detection Responses. in Proc. Euro. Conf. on Computer Vision, 2008.

[10] Breitenstein, M D., Reichlin, F., Leibe, B,. Koller-Meier E., and Gool, L. V. Robust Tracking-by-Detection using a Detector Confidence Particle Filter. in Proc. Int. Conf. Computer Vision, 2009.

[11] Xing, J., Ai, H., and Lao, L. Multi-Object Tracking thr-ough Occlusions by Local Tracklets Filtering and Global Tracklets Association with Detection Responses. Proc. Int. Conf. Computer Vision and Pattern Recognition, 2009.

[12] Comaniciu, D., Ramesh, V., and Meer, P. Kernel-based Object Tracking. IEEE Trans. PAMI, vol 25(5):565-577, 2003.

[13] Bay, H., Ess, A., Tuytelaars, T., and Gool, L.V. Speeded Up Robust Features (SURF). Computer Vision and Image Understanding, vol. 110(3): 346-359, 2008.

[14] Namboodiri, V P., Ghorawat, A., and Chaudhuri, S. Improved Kernel-Based Object Tracking Under Occluded Scenarios. Computer Vision, Graphics and Image Processing, vol.4338: 504-515, 2006.

[15] Nat. Inst. Standards and Technology (NIST), TREC Video Retrieval Evaluation, 2001-2009. [Online].Available: http://www-nlpir.nist.gov/projects/trecvid/

[16] Bernardin, K., and Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The CLEARMOT metrics. EURASIP Journal on Image and Video Processing, 2008.