MC-HOG Correlation Tracking with Saliency Proposal

Guibo Zhu[†], **Jinqiao Wang**[†], **Yi Wu**[‡], **Xiaoyu Zhang**[§], and **Hanqing Lu**[†]
[†]National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[‡]B-DAT & CICAEET, School of Information & Control,
Nanjing University of Information Science and Technology, Nanjing, 210044, Jiangsu, China
[§]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China
{gbzhu, jqwang, luhq}@nlpr.ia.ac.cn
ywu.china@yahoo.com, zhangxiaoyu@iie.ac.cn

Abstract

Designing effective feature and handling the model drift problem are two important aspects for online visual tracking. For feature representation, gradient and color features are most widely used, but how to effectively combine them for visual tracking is still an open problem. In this paper, we propose a rich feature descriptor, MC-HOG, by leveraging rich gradient information across multiple color channels or spaces. Then MC-HOG features are embedded into the correlation tracking framework to estimate the state of the target. For handling the model drift problem caused by occlusion or distracter, we propose saliency proposals as prior information to provide candidates and reduce background interference. In addition to saliency proposals, a ranking strategy is proposed to determine the importance of these proposals by exploiting the learnt appearance filter, historical preserved object samples and the distracting proposals. In this way, the proposed approach could effectively explore the color-gradient characteristics and alleviate the model drift problem. Extensive evaluations performed on the benchmark dataset show the superiority of the proposed method.

1 Introduction

Visual tracking, which is to estimate object state in an image sequence, is one of the core problems in computer vision. It has many applications, such as surveillance, action recognition and autonomous robots/car (Yilmaz, Javed, and Shah 2006; Wang et al. 2014). One robust visual tracking approach in real-world scenarios should cope with challenges as much as possible, such as occlusions, background clutter and shape deformation.

Feature representation is critical for improving the performance in object detection (Dollár et al. 2009), tracking (Henriques et al. 2015), age estimation (Li et al. 2012) and image ranking (Li et al. 2014). Gradient and color features are the most widely used ones. To be specific, Histogram of Oriented Gradient (HOG) (Dalal and Triggs 2005) features are good at describing abundant gradient information while color features like color histograms often capture rich color characteristics. For example, integral channel features proposed by (Dollár et al. 2009) and its expansions (Dollár et al. 2014) have achieved good results in object detection by concatenating gradient and color features directly. Tang et al. explored the complementary between color histogram and HOG features in a co-training framework for tracking (Tang et al. 2007). Although color and gradient features are widely used in vision based applications, there is no detailed analysis on the gradient properties for the same target in different color spaces. Therefore, it is interesting to exploit this kind of gradient properties for effective feature representation. Inspired by color naming (CN) which transforms RGB color space into an 11-D probabilistic space (Van De Weijer et al. 2009), the image pixels are projected into multiple color channels to extract gradient features for constructing a new feature descriptor: HOG extracted across Multiple Color channels (MC-HOG). It is a more natural fusion strategy than direct concatenation which needs to consider the feature normalization problem across different feature spaces.

Associated with object tracking, model drift means that the object appearance model gradually drifts away from the object due to its accumulated errors caused by online update (Matthews, Ishikawa, and Baker 2004). There are many strategies to alleviate the drift problem, *e.g.* semi-supervised learning (Grabner, Leistner, and Bischof 2008), ensemble-based learning (Tang et al. 2007; Kwon and Lee 2010), long-term detector (Kalal, Mikolajczyk, and Matas 2012) and part context learning (Zhu et al. 2015). In essence, they either explored the supervised information of the training samples or the search strategy. However, the reliability of training samples collected online is difficult to guarantee. To provide relatively less candidate regions and suppress the background interference, in this paper we introduce saliency proposals as prior information from visual saliency, which has been studied by many researchers (Itti, Koch, and Niebur 1998; Harel, Koch, and Perona 2006) and owns good characteristics for automatic target initialization and scale estimation (Seo and Milanfar 2010; Mahadevan and Vasconcelos 2013). The saliency map is taken as the prior information to obtain candidate proposals which are more efficient than exhaustive search based on sliding windows. In addition to saliency proposals, a ranking strategy is proposed to determine the importance of these proposals and estimate the

object state by exploiting the learnt appearance filter, historical preserved object samples and the distracting proposals. Therefore, the integration of saliency proposals and the ranking strategy helps a tracker to effectively update and alleviate the model drift problem.

In this paper, we propose a tracker, called as Mc-hOg Correlation sAliency tracker (MOCA), by jointly taking advantage of MC-HOG based correlation tracking and saliency proposal to explore the color-gradient characteristics and alleviate the model drift problem.

The contributions of this work are summarized as follows:

- A novel feature MC-HOG is proposed to exploit different gradient properties in various color spaces for describing the target. Extensive comparisons with other combination approaches in different color spaces show the effectiveness of MC-HOG.
- In order to reduce the risk of the model drift, saliency proposal is proposed as prior information with ranking strategy based on the learnt appearance filter, historical preserved object samples and the distracting proposals.
- Extensive experiments show that the proposed tracker achieves the state-of-the-art performance over other competitive trackers.

2 Related Work

Traditional object tracking approaches mostly focus on appearance modeling, which can be categorized roughly into generative and discriminative methods (Yilmaz, Javed, and Shah 2006). Generative methods learn an object reference model to locate the object by searching for the most similar image region, such as template matching (Matthews, Ishikawa, and Baker 2004), subspace learning (Ross et al. 2008), sparse representation (Mei et al. 2011). Although generative trackers are robust to the object occlusion and tend to obtain more accurate performance in a small searching region, they are sensitive to similar distracters in the surrounding area of the object.

In recent years, tracking-by-classification methods (Grabner, Leistner, and Bischof 2008; Babenko, Yang, and Belongie 2009; Hare, Saffari, and Torr 2011; Zhang, Ma, and Sclaroff 2014; Hong et al. 2015) have shown promising tracking performance. Many of these approaches formulate tracking as a binary classification or regression problem. The classification-based trackers (Grabner, Leistner, and Bischof 2008; Babenko, Yang, and Belongie 2009) require a set of binary labeled training instances to determine the decision boundary for distinguishing the target object from the background. The ensemble based trackers (Zhang, Ma, and Sclaroff 2014; Hong et al. 2015) proposed different ensemble strategies (*i.e.* entropy minimization (Zhang, Ma, and Sclaroff 2014) and multiple memory stores (Hong et al. 2015)) to handle the occlusion problem and achieves good performance for visual tracking. While the regression-based trackers (Bolme et al. 2010; Hare, Saffari, and Torr 2011; Henriques et al. 2012) utilize the training samples with spatial label distribution as supervised information for training better decision boundary by adopting the structured output prediction or dense sampling. Recently, the regression-based trackers have been explored and achieved good performance (Wu, Lim, and Yang 2013; Kristan et al. 2014). Especially, the series of the correlation filters-based trackers, DSST (Danelljan et al. 2014a), SAMF (Li and Zhu 2014), KCF (Henriques et al. 2015), are shown to be the top-performing trackers in accuracy in the VOT2014 challenge (Kristan et al. 2014).

Correlation filters have been investigated by many researchers in the context of visual tracking. Bolme et al. (Bolme et al. 2010) proposed an adaptive correlation filter with minimizing the output sum of squared error (MOSSE) for the target appearance in visual tracking. It can use the convolution theorem for fast learning and detection. Henriques et al. (2012) proposed circulant structure tracker (CSK) which exploited the circulant structure of adjacent subwindows for quickly learning a kernelized regularized least squares classifier of the target appearance with dense sampling. Kernelized correlation filters (KCF) (Henriques et al. 2015) is an extended version of CSK by re-interpreting the correlation filter using kernelized regression with multi-channel features. Danelljan et al. (2014b) introduced color attributes to exploit the colorful property in improving the tracking performance on colorful sequences and then proposed accurate scale estimation with one separate filter in (Danelljan et al. 2014a). Zhang et al. (2014) utilized the spatio-temporal context to interpret correlation tracking with the Bayesian framework. In a word, all of them attempt to exploit different characteristics of correlation filters for tracking, e.g., circulant structure (Henriques et al. 2012), color attributes (Danelljan et al. 2014b), kernel trick (Henriques et al. 2015), HOG conjunction with correlation tracking (Danelljan et al. 2014a; Henriques et al. 2015).

In this paper we focus on feature representation and the model drift problem in visual tracking. The aforementioned correlation trackers either rely on HOG, intensity or color features. To explore the mutual complementary between gradient and color information for enhancing the ability of feature representation, we make an extensive investigation on combination of HOG and color features in various color spaces. In addition, to handle the model drift caused by occlusion or distracters, saliency proposals are proposed as prior information and the ranking strategy for guaranteeing the correctness of the proposals, which are not considered by all aforementioned trackers to the best of our knowledge.

3 Proposed Approach

In this section, we first discuss how to extract the proposed MC-HOG in multiple color channels or color space especially in the color naming space, and then explain how to learn MC-HOG tracker with adaptive update for visual tracking. Finally, we propose to jointly utilize MC-HOG based correlation tracking and saliency proposal to alleviate the model drift problem in the tracking process.

MC-HOG Feature

Feature plays an important role in the context domain of computer vision. For example, much of the impressive progress in object detection can be attributed to the im-

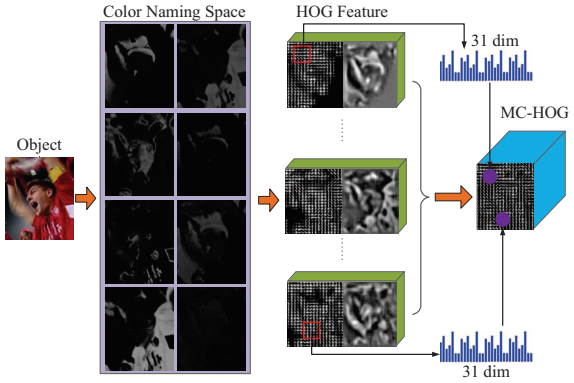

Figure 1: MC-HOG feature extraction. The HOG visualization is based on HOGgles (Vondrick et al. 2013).

provement in features, and the combination of proper features can also significantly boost the detection performance (Benenson et al. 2014). Gradient and color information are the most widely used features in object detection and tracking. Previous works (Dollár et al. 2009; Tang et al. 2007; Khan et al. 2012; 2013) have verified that, there exists a strong complementarity between gradient and color features. However, how to jointly utilize the gradient and color information for visual tracking is still an open problem.

Compared to direct concatenation, we argue that the gradient properties are different in various color spaces for the target, and therefore the extraction of gradient features from each color space is a more natural fusion strategy. Inspired by color names from linguistic view by (Berlin and Kay 1969), which contain eleven basic color terms: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow, we consider the color naming space as an example color space of our proposed feature. In computer vision, color naming is to associate RGB values with color labels which transforms RGB values into a probabilistic 11 channels color representation (Van De Weijer et al. 2009). In the paper, by investigating various color spaces, a novel visual feature, MC-HOG, is presented by calculating HOG (Felzenszwalb et al. 2010; Dollár et al. 2009) in each channel of color naming space or other color spaces and concatenate the features for all the channels as a feature descriptor. The similar operations can be easily extended to other color space as the generalized versions. For example, to balance the performance and the time complexity, we can employ MC-HOG from the Lab color space instead of 11-D color naming space.

The extraction process of MC-HOG feature is shown in Fig. 1. Firstly, the input RGB image is transformed into a color space, such as color naming space or channels. Secondly, HOG is extracted from each channel in the color space respectively. Finally, all the HOG features are concatenated in the third dimension to form a three-dimensional matrix or to a long vector. In this paper, we utilize the three-dimensional matrix representation which better fits with the correlation tracking framework.

MC-HOG Tracker with Adaptive Update

To verify the effectiveness of MC-HOG feature in visual tracking, we train a discriminative correlation filter with a training sample X with Fast Fourier Transform: $X = \mathcal{F}(\mathbf{x})$ represented by MC-HOG feature in the current frame. A desired correlation output or probability label distribution Y in the Fourier domain together with the training sample X is also used for learning the filter. Then the filter is applied to estimate the target state in the next frame. Specifically, the optimal correlation filter H is obtained by minimizing the following cost function,

$$\min_{H} \|H \circ X - Y\|_2^2 + \lambda \|H\|_2^2, \tag{1}$$

where $\circ$ is Hadamard product operator, the first term is the regression target, the second term is a L2 regularization on H, and λ controls the regularization strength. With kernel trick (Schölkopf and Smola 2002) and circulant structure (Henriques et al. 2012), kernelized correlation filters was proposed for visual tracking which allowed more flexible, non-linear regression functions integrating with multi-channel features (Henriques et al. 2015). Due to the characteristic of the kernel trick, the model optimization is still linear in the dual space even if a set of variables. Then the linear kernelized correlation filter H is represented as:

$$H = \frac{Y\Phi(X)}{K(X,X) + \lambda}, \tag{2}$$

where $\Phi(X)$ is a mapping function to compute the kernel matrix $K(.,.)$ in Fourier space.

In the process of visual tracking, the coefficients A of kernelized regularized Ridge regression and the target appearance X are updated with the following linear interpolation:

$$A = \frac{Y}{K(X,X) + \lambda}, \tag{3}$$

$$A^t = (1-\beta) * A^{t-1} + \beta * A, \tag{4}$$

$$X^t = (1-\beta) * X^{t-1} + \beta * X, \tag{5}$$

where t denotes the t-th frame and β denotes the learning rate. Actually, the update strategy works well if there is no occlusion or the object appearance changes slowly.

However, when the object is occluded, the object appearance will be updated inappropriately which may lead to the drift problem. To deal with the problem, we introduce two indicators to evaluate whether the object is occluded and adaptively adjust the learning rate. If the object is occluded, we reduce the learning rate; if else, keep the learning rate. The two indicators are Peak-to-Sidelobe Ratio (PSR) proposed by (Bolme et al. 2010) and appearance similarity. The PSR is denoted as $\frac{\delta}{g_{max}-\mu}$, where g_{max} is the maximum value of the correlation ouput and μ and δ are the mean and standard deviation of the sidelobe. The sidelobe is the rest of the pixels excluding an 11×11 window around g_{max}. We compute the appearance similarity d as follows:

$$d = exp(-\eta * ||\mathbf{x} - \mathbf{x}^{t-1}||^2), \tag{6}$$

where η is a hyperparameter which is set as 0.05, the function $||.||$ is the Euclidean distance between the object appearance $\mathbf{x}$ and $\mathbf{x}^{t-1}$, and $\mathbf{x}$ denotes the object appearance

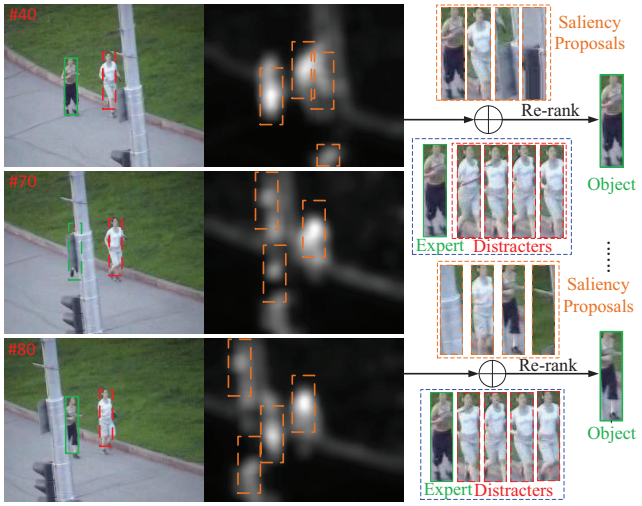

Figure 2: Object state estimation with saliency proposals and re-ranking. Best viewed in high-resolution.

in spatial domain. Compared to Eq. (5), $\mathbf{x}$ is the appearance feature by transforming the MC-HOG feature matrix of object to a vector representation. It needs to note that η is related to the appearance representation. With the PSR value and the similarity d, we adjust the learning rate β as follows:

$$\beta = \begin{cases} \gamma * \beta_{init}, & if \quad PSR \leq 30 \quad \& \quad d \leq 0.22 \\ \beta_{init}, & otherwise \end{cases} \tag{7}$$

where γ is the relative ratio to reduce the learning rate β. β_{init} is the initialization value.

For predicting the new object state, a sliding-window-like manner is necessary. Let z denotes a $M \times N \times D$ feature map extracted from an image region with size $M \times N$, D is the number of feature channels. With the convolution Theorem and circulant structure (Henriques et al. 2015), the confidence scores $S(z)$ at all locations in the image region can be computed efficiently,

$$S(z) = \mathcal{F}^{-1}\{A \circ K(X, Z)\}, \tag{8}$$

where the search region of Fourier domain $Z = \mathcal{F}(z)$, $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the discrete Fourier transform and the inverse Fast Fourier transform.

Saliency Proposal

For correlation filter-based trackers (Bolme et al. 2010; Henriques et al. 2012; 2015), there exist two main challenges: scale variation and the model drift caused by occlusion or distracter. In (Danelljan et al. 2014a), an independent scale prediction filter was presented to deal with the scale changes. A common approach to handle the model drift problem is to integrate a short-term tracker and online long-term detector, *e.g.* TLD (Kalal, Mikolajczyk, and Matas 2012). However, learning an online long-term detector relies heavily on lots of well labeled training samples which are difficult to collect. Meanwhile, the exhaustive search in whole image with sliding windows is time-consuming, especially for complex but discriminative features.

To provide relatively less proposals and suppress the background interference, in this paper we not only utilize an adaptive update strategy to learn the appearance model, but also exploit a few reliable proposals from the biologically inspired saliency map. The saliency proposals provide lots of prior information from visual saliency, which has been studied by many researchers (Itti, Koch, and Niebur 1998; Harel, Koch, and Perona 2006) and owns good characteristics for automatic target initialization and scale estimation (Seo and Milanfar 2010; Mahadevan and Vasconcelos 2013). We argue the prior information could alleviate the model drift problem caused by occlusion or distracters by providing the confident candidates and restraining the background disturbing regions.

Based on the studies for visual attention of the primate visual system (Itti, Koch, and Niebur 1998; Harel, Koch, and Perona 2006), we primarily achieve a visual saliency map and then iteratively obtain a series of candidate windows or proposals. To be specific, we first compute the visual saliency using the code from (Harel, Koch, and Perona 2006) and take the region of the last object state as the first saliency proposal; then we set the corresponding saliency value to zero and select the region with maximum saliency value as the second saliency proposal. Subsequently we further set the corresponding saliency value of the second proposal to zero, and iteratively select the saliency proposals again until the saliency value is smaller than a given threshold θ ($\theta = 0.5$). After we obtain N saliency proposals at most, we calculate the correlation output values $\mathbf{s} = \{s_1, s_2, ..., s_N\}$ with the inference process according to Eq. (8), the corresponding object centers $C = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N\}$ and the candidate object appearances $A = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N\}$ in the feature space of spatial domain.

Object State Estimation with Re-ranking

As illustrated in Fig. 2, in addition to the saliency proposals in current frame, we also preserve the historical positive object samples or experts $P = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_M\}$ and identify some hard negative samples or distracting proposals $U = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K\}$ which are supposed as distracters. M and K are the preserving sample number of positive objects and possible proposal distracters, respectively. In this paper, $M = 4$ and $K = 4$. The positive object samples are preserved every 50 frames and the distracting proposals are stored every ten frames. The second highest confident proposal in the final decision scores is identified as a distracting proposal while the highest is considered as the object state.

With the obtained proposals from historical and saliency information, we re-rank them with correlation similarity, spatial weight and ranking weight. The spatial weight is defined as a Gaussian distribution around the object position. For the i-th proposal in the t-th frame, the weight w_i is,

$$w_i = exp(-\frac{||\mathbf{c}_i^t - \mathbf{c}^{t-1}||^2}{2\sigma^2}), \tag{9}$$

where the function $||.||$ is the Euclidean distance, $\mathbf{c}^{t-1}$ denotes the predicted object center in the $(t-1)$-th frame, and

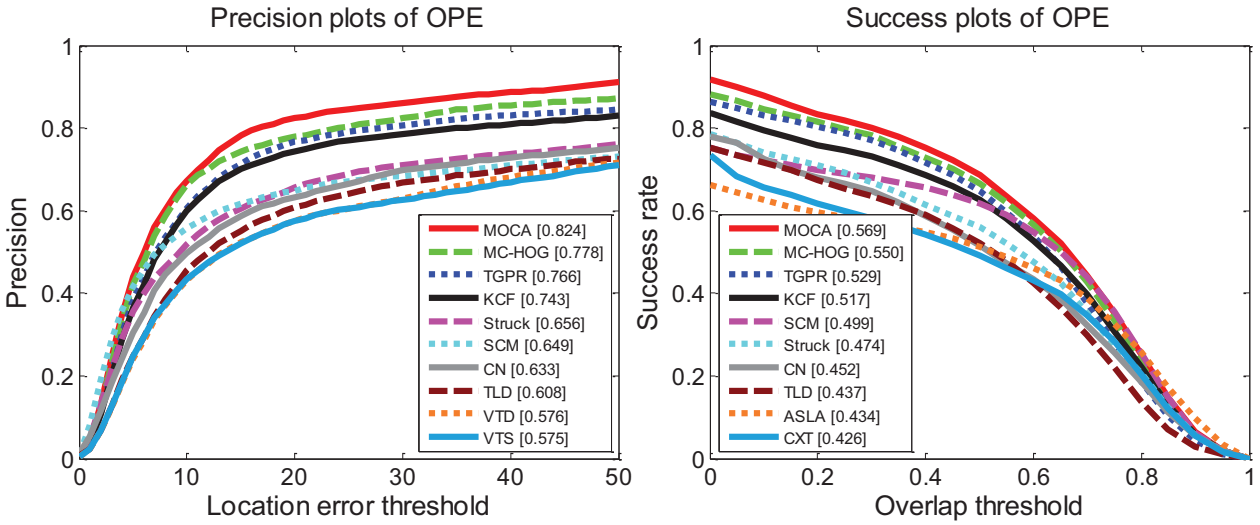

Figure 3: Precision and success plots of overall performance comparison for the 50 videos with 51 target objects in the benchmark (Wu, Lim, and Yang 2013) (best-viewed in high-resolution). The mean precision scores for each tracker are reported in the legends. Our methods are shown in *red* and *green*. In both cases our two approaches (MOCA and MC-HOG) perform favorably better than the state-of-the-art tracking methods.

$\sigma = \sqrt{w * h}$. w and h are the width and height of the search region or the cosine window in (Henriques et al. 2015).

The template similarity v_i of the i-th proposal in the current frame is computed as follows.

$$v_i = max(sim(\mathbf{a}_i, P)) - max(sim(\mathbf{a}_i, U)), \quad (10)$$

where $sim(\mathbf{a}_i, P||U)$ is the similarity values between the appearance feature of candidate $\mathbf{a_i}$ and the positive sample pool P or negative sample pool U. Based on the template similarity $\mathbf{v}$, the ranking weights $\mathbf{r} = \{r_1, r_2, ..., r_N\}$ are computed by $\mathbf{r} = exp(-\omega(Idx - 1))$, where the parameter ω is hyper-parameter, and Idx is the ranking order of the proposals by sorting the template similarity $\mathbf{v}$. We set $\omega = 0.2$. At last, we multiplies the correlation similarity, spatial weights and ranking weights to re-rank the proposals,

$$s = \max(\mathbf{s} \circ \mathbf{w} \circ \mathbf{r}), \quad (11)$$

where the corresponding position of the maximum value s is predicted as the object state in the current frame. To reduce the computational complexity, we consider the saliency proposals for re-ranking every ten frames.

4 Experiments

We evaluate our MOCA tracker on the challenging CVPR2013 Visual Tracker Benchmark (Wu, Lim, and Yang 2013), by following rigorously their evaluation protocols. There are totally 50 sequences used to evaluate the proposed approach. The experiments are performed in Matlab on an Intel Xeon 2 core 2.50 GHz CPU with $256G$ RAM.

In all the experiments, we use the *same* parameter values for all sequences (*i.e.* $\lambda = 0.0001$, $\gamma = 0.1$ and $\beta_{init} = 0.02$). We first evaluate the characteristics of MC-HOG feature in different color spaces. Then we test our proposed tracker MOCA on the benchmark dataset comparing with many competitive tracking approaches.

Experiment 1: Evaluation of MC-HOG in different color spaces

To evaluate the effectiveness of capturing the color and gradient property in different color spaces, we perform an extensive evaluation of MC-HOG in different color spaces. Although the motivation of these color features vary from photometric invariance and discriminative power to biologically inspired color representation (Danelljan et al. 2014b), we believe that the gradient properties are different in various color spaces for the target.

Table 1 shows the results of HOG extracted in different color spaces. All color representations are appropriately normalized. The conventional KCF tracker with Gray-HOG provides a mean distance precision at a threshold of 20 pixels of 74.3%. The second best results are achieved by using HOG extracted from the Luv color space with a gain of **2.5**% over the Gray-HOG while MC-HOG in color naming space achieves the best performance. As shown in Table 1, we can find different color spaces show different color and gradient properties, such as Luv and Lab are better than others except color naming, XYZ performs worst with HOG and a rich representation in 11 color channels of color naming space show a strong discriminative ability.

Experiment 2: CVPR2013 Visual Tracker Benchmark

We evaluate our methods with 10 different state-of-the-art trackers. The trackers used for comparison are: VTD (Kwon and Lee 2010), VTS (Kwon and Lee 2011), TLD (Kalal, Mikolajczyk, and Matas 2012), CXT (Dinh, Vo, and Medioni 2011), Struck (Hare, Saffari, and Torr 2011), ASLA (Jia, Lu, and Yang 2012), SCM (Zhong, Lu, and Yang 2012), CN (Danelljan et al. 2014b), KCF (Henriques et al. 2015), TPGR (Gao et al. 2014), and our trackers (MOCA and MC-

Color	Gray	RGB	Lab	Luv	YCbCr	YPbPr	YDbDr	HSV	HSI	XYZ	LMS	CN
DP1	74.3%	72.1%	75.3%	76.8%	72.0%	68.4%	71.1%	74.2%	71.3%	60.0%	61.2%	**77.8%**
DP2	91.1%	88.7%	91.1%	94%	87.7%	84.7%	87.7%	92.1%	87.4%	67.5%	67.7%	**97.4%**
M-FPS	**35.9**	20.8	18.4	19.3	22.8	20.5	22.8	20.5	20.0	20.8	18.8	16.5

Table 1: Comparison of HOG in different color spaces for tracking. The best two results are shown in **red** and blue. The results are presented using both mean distance precision (DP1) and median distance precision (DP2) over all 50 sequences (Wu, Lim, and Yang 2013). While the sequence is gray, we only adopt the conventional intensity channel for HOG extraction. In both cases, the best results are obtained by using the MC-HOG feature. M-FPS: mean frames per second.

HOG), etc. The overall performance is shown in Fig. 3. The public codes of the comparative trackers are provided by the authors and the parameters are fine tuned. All algorithms are compared in terms of the initial positions in the first frame from (Wu, Lim, and Yang 2013). Their results are also provided with the benchmark evaluation (Wu, Lim, and Yang 2013) except KCF, CN, and TGPR. Here, KCF used HOG feature and the gaussian kernel which achieved the best performance in (Henriques et al. 2015). CN's source code was originated from (Danelljan et al. 2014b). It was modified to adopt the raw pixel features as (Henriques et al. 2015) while handling the grey-scale images.

Fig. 3 shows precision and success plots which contains the mean distance and overlap precision over all the 50 sequences. The values in the legend are the mean precision score and AUC, respectively. Only the top 10 trackers are displayed for clarity. Our approaches MOCA and MC-HOG both improve the baseline HOG-based KCF tracker with a relative reduction in accuracy. Moreover, our MC-HOG tracker improves the precision rate of the baseline method KCF from 74.3% to **77.8**%, and then MOCA boosts the MC-HOG tracker with a relative gain of **4.6**%. Moreover, our MC-HOG and MOCA trackers improve the success rate of their baseline methods from 51.7% to **55.0**%, and from 55.0% to **56.9**%. In (Henriques et al. 2015), the performance of KCF is better than Struck in precision of predicting the object state. Overall, our trackers are better than the other trackers and achieves a significant gain. Although our method does not estimate scale variations, it still provides encouraging results compared other competitive trackers in mean overlapping precision.

Attribute-based Evaluation: We perform a comparison with other methods on the 50 sequences respect to the 11 annotated attributes (Wu, Lim, and Yang 2013). Fig. 4 shows some example precision plots of four attributes. For occlusion, out of view or background clutter sequences, MOCA is much better than MC-HOG because of saliency proposals and the ranking strategy. Saliency proposals provide proper candidates and suppress the background interference for the subsequent re-ranking process as illustrated in Fig. 2. Because the ranking strategy explores and exploits the learnt appearance filter, motion penalization, the historical object experts, and the distracting proposals. The learnt appearance filter and motion penalization can handle the object appearance changes. The historical object experts can verify the correctness of the object candidates while reserving the distracting proposals suppresses the distracting regions. Both of them can alleviate the drift problem caused by occlusion

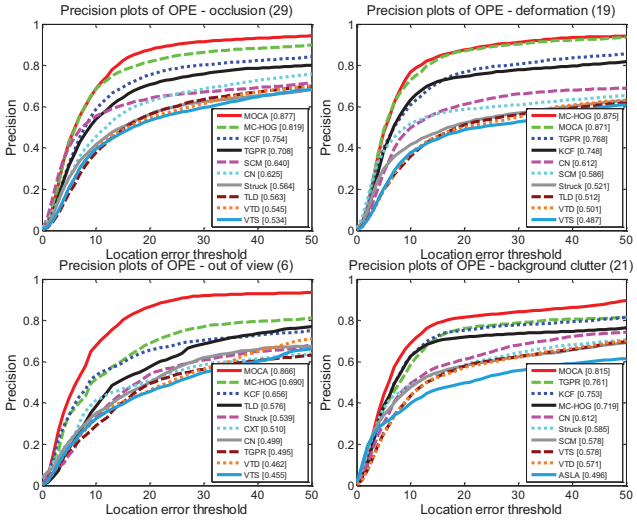

Figure 4: Precision plot for sequences with attributes: occlusion, non-rigid deformation, out-of-view and background clutter. The proposed MOCA tracker is the most resilient to all of nuisances. Best viewed in high-resolution.

or distracters. For deformation sequences, MOCA and MC-HOG also provide superior results compared to other existing methods. This is due to the fact that color attributes possess a certain degree of photometric invariance while preserving discriminative power.

5 Conclusion

In this paper, we have developed MC-HOG correlation tracking with saliency proposals and a ranking strategy. Our experimental results demonstrate the complementary of different color spaces and gradient features, and show that exploiting different gradient properties in various color spaces for the target is helpful for the tracking performance. Moreover, we have showed that the MOCA tracker by jointly utilizing MC-HOG based correlation tracking and saliency proposals with the ranking strategy can also alleviate the model drift problem caused by occlusion or distracters. Finally, extensive experiments show that our tracker outperforms the state-of-the-art methods on the benchmark dataset.

6 Acknowledgments

This work was supported by 863 Program (Grant No. 2014AA015104), and National Natural Science Founda-

tion of China (Grant No. 61273034, 61332016, 61370036, 61501457).

References

Babenko, B.; Yang, M. H.; and Belongie, S. 2009. Visual tracking with online multiple instance learning. In *CVPR*, 983–990. IEEE.

Benenson, R.; Omran, M.; Hosang, J.; and Schiele, B. 2014. Ten years of pedestrian detection, what have we learned? In *Workshop of ECCV*, 613–627. Springer.

Berlin, B., and Kay, P. 1969. *Basic color terms: their universality and evolution*. Berkeley and Los Angeles: University of California Press.

Bolme, D.; Beveridge, J.; Draper, B.; and Lui, Y. 2010. Visual object tracking using adaptive correlation filters. In *CVPR*, 2544–2550. IEEE.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, 886–893. IEEE.

Danelljan, M.; Häger, G.; Khan, F.; and Felsberg, M. 2014a. Accurate scale estimation for robust visual tracking. In *BMVC*.

Danelljan, M.; Shahbaz Khan, F.; Felsberg, M.; and Van de Weijer, J. 2014b. Adaptive color attributes for real-time visual tracking. In *CVPR*. IEEE.

Dinh, T.; Vo, N.; and Medioni, G. 2011. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR*, 1177–1184. IEEE.

Dollár, P.; Tu, Z.; Perona, P.; and Belongie, S. 2009. Integral channel features. In *BMVC*, 5.

Dollár, P.; Appel, R.; Belongie, S.; and Perona, P. 2014. Fast feature pyramids for object detection. *TPAMI* 36(8):1532–1545.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part based models. *TPAMI* 32(9):1627–1645.

Gao, J.; Ling, H.; Hu, W.; and Xing, J. 2014. Transfer learning based visual tracking with gaussian processes regression. In *ECCV*. Springer. 188–203.

Grabner, H.; Leistner, C.; and Bischof, H. 2008. Semi-supervised on-line boosting for robust tracking. In *ECCV*. Springer. 234–247.

Hare, S.; Saffari, A.; and Torr, P. 2011. Struck: Structured output tracking with kernels. In *ICCV*, 263–270. IEEE.

Harel, J.; Koch, C.; and Perona, P. 2006. Graph-based visual saliency. In *NIPS*, 545–552.

Henriques, J.; Caseiro, R.; Martins, P.; and Batista, J. 2012. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*. Springer. 702–715.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2015. High-speed tracking with kernelized correlation filters. *TPAMI* 37(3):583–596.

Hong, Z.; Chen, Z.; Wang, C.; Mei, X.; Prokhorov, D.; and Tao, D. 2015. Multi-store tracker (muster): a cognitive psychology inspired approach to object tracking. In *CVPR*, 749–758.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *TPAMI* 20(11):1254–1259.

Jia, X.; Lu, H.; and Yang, M. 2012. Visual tracking via adaptive structural local sparse appearance model. In *CVPR*, 1822–1829. IEEE.

Kalal, Z.; Mikolajczyk, K.; and Matas, J. 2012. Tracking-learning-detection. *TPAMI* 34(7):1409–1422.

Khan, F.; Anwer, R.; van de Weijer, J.; Bagdanov, A.; Vanrell, M.; and Lopez, A. 2012. Color attributes for object detection. In *CVPR*, 3306–3313. IEEE.

Khan, F.; Anwer, R.; van de Weijer, J.; Bagdanov, A.; Lopez, A.; and Felsberg, M. 2013. Coloring action recognition in still images. *IJCV* 105(3):205–221.

Kristan, M.; Pflugfelder, R.; Leonardis, A.; Matas, J.; Porikli, F.; Cehovin, L.; Nebehay, G.; Fernandez, G.; Vojir, T.; Gatt, A.; et al. 2014. The visual object tracking vot2014 challenge results. In *Workshop of ECCV*. springer.

Kwon, J., and Lee, K. 2010. Visual tracking decomposition. In *CVPR*, 1269–1276. IEEE.

Kwon, J., and Lee, K. 2011. Tracking by sampling trackers. In *ICCV*, 1195–1202. IEEE.

Li, Y., and Zhu, J. 2014. A scale adaptive kernel correlation filter tracker with feature integration. In *Workshop of ECCV*. Springer.

Li, C.; Liu, Q.; Liu, J.; and Lu, H. 2012. Learning ordinal discriminative features for age estimation. In *CVPR*, 2570–2577. IEEE.

Li, C.; Liu, Q.; Liu, j.; and Lu, H. 2014. Ordinal distance metric learning for image ranking. *TNNLS* 1551 – 1559.

Mahadevan, V., and Vasconcelos, N. 2013. Biologically inspired object tracking using center-surround saliency mechanisms. *TPAMI* 35(3):541–554.

Matthews, I.; Ishikawa, T.; and Baker, S. 2004. The template update problem. *TPAMI* 26(6):810–815.

Mei, X.; Ling, H.; Wu, Y.; Blasch, E.; and Bai, L. 2011. Minimum error bounded efficient l1 tracker with occlusion detection. In *CVPR*, 1257–1264. IEEE.

Ross, D.; Lim, J.; Lin, R.; and Yang, M. 2008. Incremental learning for robust visual tracking. *IJCV* 77(1-3):125–141.

Schölkopf, B., and Smola, A. 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Seo, H., and Milanfar, P. 2010. Visual saliency for automatic target detection, boundary detection, and image quality assessment. In *ICASSP*, 5578–5581. IEEE.

Tang, F.; Brennan, S.; Zhao, Q.; and Tao, H. 2007. Co-tracking using semi-supervised support vector machines. In *ICCV*, 1–8. IEEE.

Van De Weijer, J.; Schmid, C.; Verbeek, J.; and Larlus, D. 2009. Learning color names for real-world applications. *TIP* 18(7):1512–1523.

Vondrick, C.; Khosla, A.; Malisiewicz, T.; and Torralba, A. 2013. Hoggles: Visualizing object detection features. In *ICCV*, 1–8. IEEE.

Wang, J.; Fu, W.; Liu, J.; and Lu, H. 2014. Spatiotemporal group context for pedestrian counting. *TCSVT* 24(9):1620–1630.

Wu, Y.; Lim, J.; and Yang, M. H. 2013. Online object tracking: A benchmark. In *CVPR*, 2411–2418. IEEE.

Yilmaz, A.; Javed, O.; and Shah, M. 2006. Object tracking: A survey. *CSUR* 38(4):13.

Zhang, J.; Ma, S.; and Sclaroff, S. 2014. Meem: Robust tracking via multiple experts using entropy minimization. In *ECCV*. Springer. 188–203.

Zhong, W.; Lu, H.; and Yang, M. 2012. Robust object tracking via sparsity-based collaborative model. In *CVPR*, 1838–1845. IEEE.

Zhu, G.; Wang, J.; Zhao, C.; and Lu, H. 2015. Weighted part context learning for visual tracking. *TIP* 24(12):5140–5151.