Multiple Hypothesis Tracking with Sign Language Hand Motion Constraints

Mark Borg(✉) and Kenneth P. Camilleri

Systems and Control Engineering, Faculty of Engineering,
University of Malta, Msida, Malta
mborg2005@gmail.com

Abstract. In this paper, we propose to incorporate *prior* knowledge from sign language linguistic models about the motion of the hands within a multiple hypothesis tracking framework. A critical component for automated visual sign language recognition is the tracking of the signer's hands, especially when faced with frequent and persistent occlusions and complex hand interactions. Hand motion constraints identified by sign language phonological models, such as the hand symmetry condition, are used as part of the data association process. Initial experimental results show the validity of the proposed approach.

Keywords: Sign langage recognition · MHT · Tracking

1 Introduction

Hand tracking is a critical component of vision-based automated sign language recognition (ASLR) systems [1], as the hands constitute the main articulators for signing. The position of the hands, their motion, and the shapes that the hands take, all are discriminative linguistic features that contribute to the semantic meaning in sign recognition.

Object tracking is a very challenging problem, mostly due to the noisy, compressed nature of videos, the presence of motion blur, the loss of depth information, and the high variability in illumination and scene conditions. In multi-object tracking, the interaction between objects and occlusion events, make consistent labelling of objects across video frames an especially hard problem.

For ASLR in particular, tracking faces problems of frequent and persistent hand and face occlusions, and complex hand motions and interactions (like crossovers and bounce-back events), since signing occurs within a small volume of space centred on the signer. The non-rigid and articulated nature of the hands gives rise to large variations in pose and appearance, as well as issues of self-occlusion and self-shadowing. Keeping track of both (unadorned) hands over long video sequences is a very challenging problem, often fraught with the loss of hand identity and mismatch errors, which can severely degrade the performance of subsequent sign recognition modules. Thus, the core issue in multi-object tracking, like ASLR, is *data association*, i.e., determining which acquired observation corresponds to which of the objects being tracked [13].

© Springer International Publishing Switzerland 2015
G. Azzopardi and N. Petkov (Eds.): CAIP 2015, Part II, LNCS 9257, pp. 207–219, 2015.
DOI: 10.1007/978-3-319-23117-4_18

Earlier works in tracking adopted a deterministic approach for data association, where the correspondences depend only on the preceding and current video frame observations – thus termed f2f tracking. These methods generally define a cost function for associating each object at time $t-1$ to observations at time t, based on some motion constraints (e.g. proximity) and/or similarity measures (e.g., appearance or shape) [10]. Minimisation of the cost function is then formulated as a combinatorial optimisation problem, which can be restricted to 1-to-1 associations only, thus allowing for the use of optimal linear assignment algorithms such as the Kuhn-Munkres Hungarian algorithm [13]. In [6] a voting algorithm is used for data association, while [12] adopts a global nearest neighbour rule-based approach. These deterministic approaches are forced to make a *hard association decision* in each video frame, and thus a single incorrect association at a particular point in time affects all of the subsequent tracking – they can't recover from object label switches, and loss of object IDs. Later works, like that of [7], extend the data association process to multi-frame association, i.e., finding object and observation correspondences over a set of consecutive frames. This allows for the application of more constraints on temporal and spatial coherency. The data association problem now becomes a graph theoretic problem, i.e., finding the best unique path for each object within the set of frames, offering a degree of robustness against occlusion events that are shorter in duration than the temporal window used for the data association.

In contrast to f2f and multi-frame methods, statistical approaches to multi-object tracking like JPDAF (joint probabilistic data association filter) [3], use *soft association decisions*, whereby the tracked object is associated with all the feasible observations that it can be matched to, and is updated via a weighted combination of these observations. The multiple hypothesis tracking (MHT) algorithm [4,9] adopts a different strategy than the single-hypothesis tracking methods discussed earlier that only keep a single hypothesis about the past. Instead, the MHT algorithm employs a deferred decision-taking mechanism for data association, by keeping multiple hypotheses about the past, and then propagating these hypotheses into the future in anticipation that subsequent data will resolve the uncertainty about which of the multiple hypotheses is the correct one. Because of the exponential increase in the number of hypotheses created by the MHT algorithm, pruning is required – this is accomplished by a sliding temporal window, as well as by discarding low-probability hypotheses. In [1], MHT is used within an ASLR context for hand tracking in the presence of skin segmentation errors. They also make use of an anatomical hand model to eliminate anatomically impossible hypotheses. Other tracking approaches include: tracklet-based tracking, where detection of tracklets is followed by the subsequent linking of the tracklets into longer tracks [5]; Bayesian network based tracking [8]; tracking based on random finite sets. A review of tracking methods is found in [10,11].

In this paper, we propose an MHT-based framework for our ASLR system, that incorporates *prior* knowledge about the constraints on hand motion as described by sign language linguistic models. We believe that the use of this knowledge yields

an improvement in the tracking performance of MHT, especially when the tracker is dealing with complex hand interactions and occlusion events.

Sign language phonological models identify a number of constraints about hand motion, and the position of one hand in the signing space in relation to the other one. These constraints could be exploited by an MHT-based tracker in order to reduce the space of possible hypotheses. For example, the "symmetry" and "dominance" conditions limit the role of the non-dominant hand to serve as either a duplicate articulator (giving rise to the so-called "h2-S" signs), or as a place of articulation for the dominant hand (so-called "h2-P" signs). In "h2-S" signs, the articulation of h2 (the non-dominant hand) is symmetric to that of h1 (the dominant hand), and both must have the same handshape; while in "h2-P" signs, h2 is stationary and h1 moves using h2 as a place of reference against which the motion is performed [2]. Examples of "h2-P" and "h2-S" signs can be seen in Figures 2 & 3 respectively.

The main contribution of our work is the proposed integration of these hand motion constraints within the probabilistic framework of MHT. In particular, these constraints are incorporated within the hypothesis evaluation equation of the MHT algorithm via the use of probabilistic density maps. We demonstrate our approach by implementing one of these constraints, mainly the symmetry constraint "h2-S". Since symmetric hand motions can sometimes suffer from out-of-sync issues, our proposal takes this into account. While the MHT algorithm has been used within the context of ASLR, for example in [1], to the best of our knowledge there are no works that implement what we propose here.

Our use of the hand symmetry constraint, bears some resemblance in idea to the "common motion constraint" as described by [14,15]. In [14] the special problem of multi-target tracking is discussed, where a group of targets are highly correlated in their motion, usually exhibiting a common motion pattern with some individual variations; e.g., dancing cheerleaders.

The rest of this paper is organised as follows: Section 2 gives an overview of our ASLR system and tracking framework; Section 3 outlines the MHT algorithm which forms the basis of our tracking framework; Section 4 describes our proposed approach; Section 5 reports initial experimental results; We conclude the paper in Section 6 and highlight future work.

2 Overview of Our System

The multiple hypotheses based tracking framework proposed in our paper forms part of an ASLR system. The work described here concentrated on (1) object detection, (2) tracking, and (3) the incorporation of sign language hand motion constraints within the tracking process. The object detection stage locates the face, computes skin and motion likelihoods, which in turn serve for detecting the hands – these provide the "observations" that are fed to the second stage. The tracking stage is made up of a number of steps: track prediction, gating, hypotheses formation about the associations of observations to tracks, followed by the adjustment of hypotheses' likelihoods, and their evaluation and eventual pruning. The adjustment of the likelihoods makes use of sign language

hand motion constraints, and constitutes the main contribution of this paper. Additional information can also be incorporated into this step, like kinematic hand/upper body models and contextual information, via the use of the same mechanism and principles – this is described later on as future work. Finally, the most likely hypothesis about the position and motion of the hands is fed to the sign recognition module of our system, which is still works in progress.

3 The MHT Algorithm

Reid [9] was the first to describe MHT using a strong mathematical formulation. As mentioned in the introduction, in contrast to single-hypothesis tracking methods that make a hard decision in each time step as regards to which observations are associated with which targets, the idea behind the MHT algorithm is to generate all possible association hypotheses at any one time step and then rely on future information to resolve any ambiguities and to select the most probable hypothesis amongst them. The hypothesis generation stage implicitly caters for various observation-to-track association scenarios, such as new track initiation and termination (e.g., targets entering/leaving the camera's FOV), targets which are unobserved for some time, and observations arising from noise (false alarms). As a new set of observations arrives with each new time step, newly-generated hypotheses are added to the previous ones, thus forming a tree structure.

To avoid a combinatorial explosion in the number of hypotheses generated and maintained by the algorithm, a number of pruning techniques are applied to make the tracking more tractable – these include pruning low probability hypotheses, specifying a maximum number of hypothesis, N-scan pruning, and applying clustering techniques [4].

Hypotheses clustering helps to divide the tracking problem into independent and smaller sub-problems which can be solved separately. First, observations are gated with existing tracks; any observations falling outside the validation gate of a track are considered to have a zero probability of being associated with that track, and thus can be safely ignored (assuming the statistical model used to obtain the validation gate is valid). This helps to remove those observation-to-track associations which are considered to be physically impossible. Grouping collections of tracks linked by observations then gives rise to clusters of hypotheses, which in turn results in spatially disjoint hypothesis trees [16]. In N-scan pruning, a sliding temporal window is applied to the tree of hypotheses. Within this window, multiple hypotheses are maintained and propagated in time as the window slides forward. But at the rear end of the window, a hard decision is made on the most likely hypothesis taking into consideration future observational evidence present in the window. Thus the depth of the hypotheses trees are limited to be at most equal to the window size. In [4,17] it was shown that MHT can achieve good tracking performance with quite shallow tree depths.

In [17], the efficiency of the MHT is greatly improved via the use of Murty's algorithm, which addresses the main inefficiency of the original MHT – mainly that a lot of computation is wasted on the generation and propagation of many

hypotheses which in the end are discarded and never used. This is achieved by finding the m-best associations hypotheses in the current time step instead of using all possibilities. More detail on the MHT algorithm can be found in [3,4].

4 MHT and Sign Language Constraints

In this section we will describe our MHT-based system, the features we have selected for detecting the objects to be tracked (observation acquisition), and the chosen target representation. Then we will describe the proposed incorporation of sign language hand motion constraints within our tracking process.

4.1 Features for Object Detection

Choosing the right *features* to be used in a multi-object tracking system is an extremely important task – many times, the choice of features depends on the tracking domain in question. Ideally the chosen features should be unique, in order to facilitate the detection of the objects to be tracked (the observations), and should be computationally efficient to extract [10]. We use motion-based features in our ASLR tracking system. To make the extraction of these features as efficient as possible, we apply pre-filtering based on skin colour and frame differencing. We also employ a face detector, both for face localisation purposes, as well as for performing system initialisation, such as that of learning the skin colour model. A tracking by detection approach is adopted for face localisation, via the use of the Viola-Jones face detector [26]. The assumption behind our face localisation approach is that the face of the signer is frontal (or near-frontal) with respect to the camera's viewpoint. Once the face of the signer is detected, a body-centred coordinate system is defined and scaled according to the size of the face. To improve the accuracy of face localisation, a constant-velocity Kalman filter is used to smooth out the noise in global head motion.

We employ an adaptive skin colour classifier [18] for generating the skin likelihood map. The skin model used by this classifier is initialised via face detection as follows: a 24×24 mask, generated off-line using several hundred images of different persons, is applied to the face region that is found by the face detector – this mask indicates which pixels within the face region are most likely to be skin; then working within the normalised RGB colour space, a parametric skin colour model is estimated.

A motion likelihood map is generated via a weighted frame differencing algorithm. Combined together, the skin likelihood and motion likelihood maps, serve as a fast pre-processing stage by filtering out most areas of the image (moving skin-coloured regions are expected to be small in size and number). Thus we avoid having to run the costlier feature extraction process over the full image.

The motion-based features used in our ASLR system consist of clusters of KLT features (corners) exhibiting a similar affine motion model. KLT features are first located within moving skin regions (as filtered by the skin and motion

likelihood maps) using the method described in [19] based on a goodness-to-track quality metric. The motion information of the chosen KLT features are then obtained via a multi-scale sparse optical flow algorithm. Similar to the work of [12], and relying on the notion of "common motion constraint", we apply an iterative method for clustering KLT features by their affine motions. The RANSAC scheme is used for robust affine motion model fitting, because of its high breakdown point (can tolerate up to 50% outliers). Finally, the clusters of KLT features and their associated affine motion models constitute the *observations* that will be fed to the MHT stage of our tracking system

$$z_i^t = \left\{ \{k_j\}_{\forall j \in C_i}, \mathbf{A}_i^t \right\} \tag{1}$$

where k_j is the j^{th} KLT feature, C_i is the i^{th} cluster of KLT features, and A_i^t is the affine motion model fitted to the KLT features $\{k_j\}$ of the i^{th} cluster

$$\mathbf{A}_i^t = \begin{bmatrix} a_0 & a_1 & a_2 \\ a_3 & a_4 & a_5 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

4.2 Target Representation

Many different representations have been used in the object tracking literature such as: appearance-based representations (templates, active appearance models, etc.), motion-based representations, and shape-based representations (silhouettes, contours, primitive geometric shapes, articulated shape models, skeletal models, etc.) [10]. We represent objects (the hands and face) by their affine motion model $\mathbf{A}$, centroid (x, y), and their spatial extent (bounding box having width w and height h). Assuming a linear dynamic process, a constant-velocity Kalman filter is used with the following state $\mathbf{x}$ and state transition matrix $\mathbf{F}$, where $\mathbf{I}_{16\times16}$ is an identity sub-matrix

$$\mathbf{x} = \left[x, y, \frac{w}{2}, \frac{h}{2}, a_0, a_1, \cdots, a_5, \dot{x}, \dot{y}, \frac{\dot{w}}{2}, \frac{\dot{h}}{2}, \dot{a}_0, \dot{a}_1, \cdots, \dot{a}_5 \right]^T$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{I}_{16\times16} & \mathbf{I}_{16\times16} \cdot dt \\ \mathbf{0}_{16\times16} & \mathbf{I}_{16\times16} \end{bmatrix} \tag{3}$$

This Kalman filter allows us to predict the position of the hands $\bar{x}_j^t$ at time t. The predictions of the width $w/2$ and height $h/2$ at time t given by the KF are used for occlusion prediction by checking for bounding box overlap of the 2 hands – this information is used to set the occlusion terms of the MHT. The affine motion terms a_0 to a_5 in the KF are for smoothing the target's affine motion model $\mathbf{A}$, used for KLT feature clustering and replenishment (see Section 4.1).

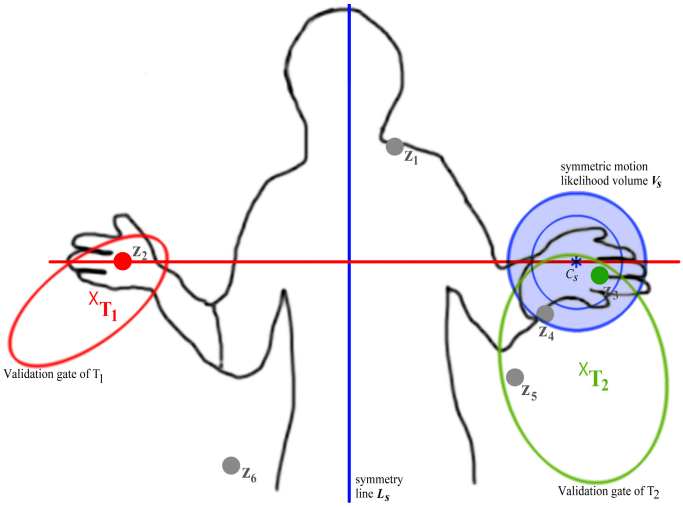

Fig. 1. Use of the symmetric motion constraint in tracking

4.3 Tracking with Sign Language Hand Motion Constraints

Our proposal in this paper is to incorporate constraints on hand motion based on sign language linguistic models. In particular, we concentrate on one such constraint from the sign language literature called the "symmetry" condition ("h2-S") [2]. This states that for the majority of 2-handed signs, the non-dominant hand serves as a duplicate articulator, i.e., its movement mirrors that of the dominant hand. Figure 1 above illustrates in a schematic way, how we integrate the symmetric motion constraint within the tracking process. Our tracking process assumes that the 2 hands are the main moving objects in the scene, and that the right and left hands start on the right and left side of the body respectively. Using the face detection results for time t, we first identify the body-centred line of symmetry L_s. Currently we ignore any sideways leaning of the signer's upper body and keep L_s oriented vertically. Then during the MHT hypothesis generation stage, given an association hypothesis ψ_i^t that associates object T_1 with observation z_2, we locate the corresponding point C_s reflected by the line of symmetry L_s, and define the symmetric motion likelihood volume V_s, depicted in Figure 1. Currently we adopt a non-parametric approach for the pdf of V_s, i.e., using a probability density map. For object T_2, the probability of associating it with the set of observations in its validation gate is then computed via a weighted combination of the standard MHT's observation-target probability ([3,4]) and the probability given by our density map:

$$\mathcal{N}'\left(z_i^t\right) \triangleq \alpha \cdot \mathcal{N}\left(z_i^t | \bar{x}_j^t, \Sigma_{ij}^t\right) + (1-\alpha) \cdot \mathbf{M}_t^{V_s} \tag{4}$$

where $\mathbf{M}_t^{V_s}$ is the probability density map of V_s at time t, and α is a weighting factor. Continuing with the example depicted in Figure 1, the hypothesis that object T_2 is associated with observation z_3 now has a stronger bias, because of

the addition of $\mathbf{M}_t^{V_s}$ term. An issue that our proposed incorporation of symmetric motion has to contend with is when the 2 hands are moving slightly out of sync – one hand lags behind the other in its mirrored trajectory position. We solve this issue by adopting a temporal window approach for the update of the probability density map $\mathbf{M}_t^{V_s}$ with a forgetting mechanism with rate γ

$$\left(\mathbf{M}_t^{V_s}\right)' = \gamma\left(\mathbf{M}_t^{V_s}\right) + (1-\gamma)\left(\mathbf{M}_{t-1}^{V_s}\right) \tag{5}$$

5 Experiments

In order to evaluate the effectiveness of our approach, video sequences from the ECHO Sign Language (NGT) Corpus [21] are used. These are colour sequences, with 352×288 resolution, running at 25 fps, and taken with a fixed camera. Although signing occurs within a simplified environment (indoors, constant illumination, plain background, signer wearing dark clothes), these sequences exhibit frequent occlusions of the hands and the face, often of medium to long duration, and with lots of complex hand interaction ('bounce backs', 'crossovers') events. Thus we believe that videos from this corpus constitute a good test bed for our proposed tracking system. As no tracking-related ground truthing is available for these video sequences, the ViPER-GT toolkit [20] was used to generate ground truth of the face and the hands for around 6000 video frames – via the visual annotation feature of ViPER-GT of the positions and the bounding boxes.

5.1 Experimental Setup

Our system was implemented in C++ and makes use of the OpenCV library. Our MHT implementation is based on the original MHT library reported in [17]; but it also includes additional modifications as suggested in [24], mainly for computational efficiency reasons, and the addition of the occlusion terms as described in [22]. Our MHT implementation can benefit from further improvements, such as more use of parallelism, especially for handling the disjoint hypothesis clusters – future work will address this.

Several algorithms used in our system have a number of configurable parameters – many of these values were set empirically. For the MHT algorithm, the temporal sliding window (used for N-scan pruning) was set to 25 frames (1 second) – this is adequate as the majority of signs last less than 1 second, thus achieving a balance between the size of the mantained hypothesis tree for sign recognition accuracy and the real-time execution speed requirement. The *a priori* probability values for track detection p_{det}, track termination p_{term}, and occlusion events p_{occl}, were configured as 0.7, 0.1 and 0.2 respectively (subject to the condition $p_{det} + p_{term} + p_{occl} = 1$). And the Poisson expectations for false alarms and new tracks were set to: $\lambda_{fa} = 3$, and $\lambda_{new} = 1$, respectively. Even though the number of objects being tracked in an ASLR context is fixed (2 hands and a face), the parameter values for new tracks (λ_{new}) and track termination (p_{term}) must cater for the potential loss and recovery of the targets of interest, as well

as handling cases where multiple observations are returned for a single target (e.g., when apart from the hand, the arm is also visible).

Three experiments were performed. For the first experiment, we used a standard frame-to-frame (single hypothesis) tracking system and ran it on the mentioned video sequences – this system, referred to as 'baseline F2F' here, serves as a lower performance bound against which MHT-based tracking will be compared. The 'baseline F2F' system makes use of the same object detection and representation (KLT features) stage as the proposed system, but instead uses global nearest neighbour (via the Kuhn-Munkres Hungarian algorithm) for data association and a simplified rule-based track maintenance system. The second experiment made use of our MHT-based tracking system, but without the incorporation of the sign language hand motion constraints, i.e., using a regular MHT algorithm. The third and final experiment is the one making use of our contribution described in this paper – referred to as 'MHT+SL constraints' here.

The three experiments were executed on a 2.0 GHz PC with an Intel dual core CPU, and 16Gb of RAM. Tracking executed in real-time (at 20 to 23 fps).

5.2 Evaluation of Tracking Results

For evaluation purposes, we adopted the CLEAR metrics MOTP and MOTA [23]. MOTP (multiple object tracking precision) measures how well the positions of the hands are estimated by the tracker

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\Sigma_t c_t} \tag{6}$$

where $d_{i,t}$ is a distance score between the ground truth object $g_{i,t}$ and its corresponding tracker output; c_t is the number of successfully tracked objects at time t. In our evaluation we chose the overlap ratio between the ground truth's bounding box and that of the tracker output as the distance score $d_{i,t} = \frac{|g_{i,t} \cap o_{i,t}|}{|g_{i,t} \cup o_{i,t}|}$ MOTP score values are in range [0..1], with 0 indicating a perfect match.

MOTA (multiple object tracking accuracy) measures the number of mistakes that the tracker makes in terms of missed object detections (false negatives, FNs), false positives (FPs), and the number object mismatches (object label/identity switches, MMEs) that occur.

$$MOTA = 1 - \left[\frac{\sum_t \{FN_t + FP_t + MME_t\}}{\sum_t g_t} \right] \tag{7}$$

Thus MOTA gives an indication of the tracker's performance at keeping accurate trajectories, independent of the tracker's precision in estimating object positions. MOTA score values can range from negative values to 1.0 (perfect accuracy).

5.3 Discussion

The results of the quantitative evaluation using the CLEAR metrics are given in Table 1 on the next page. Also shown are the normalised FN, FP, and

Table 1. Comparative performance of MHT with SL constraints

Method	MOTP	MOTA	$FN/\sum_t g_t$	$FP/\sum_t g_t$	$MME/\sum_t g_t$
baseline F2F	0.550	0.022	0.298	0.371	0.309
MHT	0.603	0.281	0.236	0.325	0.158
MHT+SL constraints	0.601	0.313	0.225	0.328	0.134

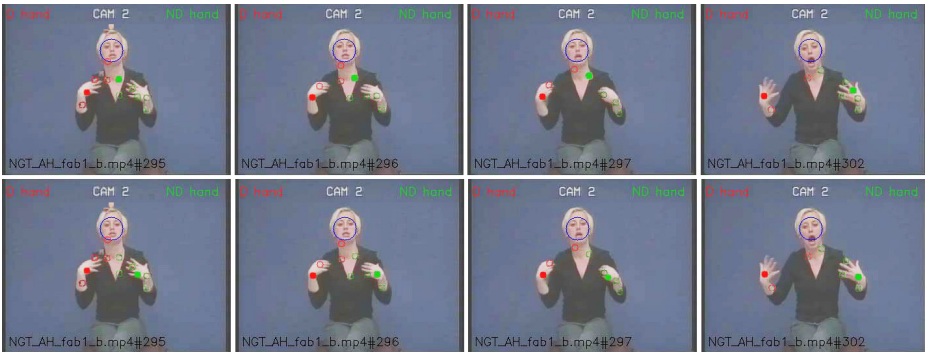

Fig. 2. Tracking results of regular MHT (top) and "MHT+SL constraints" (bottom). Correct tracks of the dominant hand (filled red circles) and non-dominant hand (filled green circles), while pruned hypotheses are shown as open circles.

mismatch rates. As can be seen, our proposed approach has an overall better performance than regular MHT. While there is only a marginal improvement in tracking precision (MOTP), tracking accuracy (MOTA) exhibits a more evident improvement. The factor that contributes most to this improvement in accuracy is the reduced number of object label mismatches (MMEs). In other words, the inclusion of constraints on hand motion based on sign language linguistic models increases the robustness of the tracking system to identity switches of the 2 hands. Figures 2 & 3 on the facing page illustrate this qualitatively.

The top row in Figure 2 gives the results of the regular MHT, while the bottom row shows the results of our approach. Prior to the first video frame shown, the hands were partially occluding each other near the neck area. Upon emerging from the occlusion event, the non-dominant hand in the regular MHT, is incorrectly matched to spurious observations (caused by shadows on the neck created by head motion). In our approach, the correct tracking of the dominant hand plus the rule on symmetric hand motion, help to increase the likelihood of the correct hypothesis of the non-dominant hand. Thus, with support from the dominant hand, the non-dominant hand is tracked successfully throughout all the frames – in contrast, the regular MHT only recovers successfully from the occlusion event in the last video frame of Figure 2.

In Figure 3, more results from our proposed tracking system, in the presence of complex hand interactions and cross-over events, are shown. Our approach

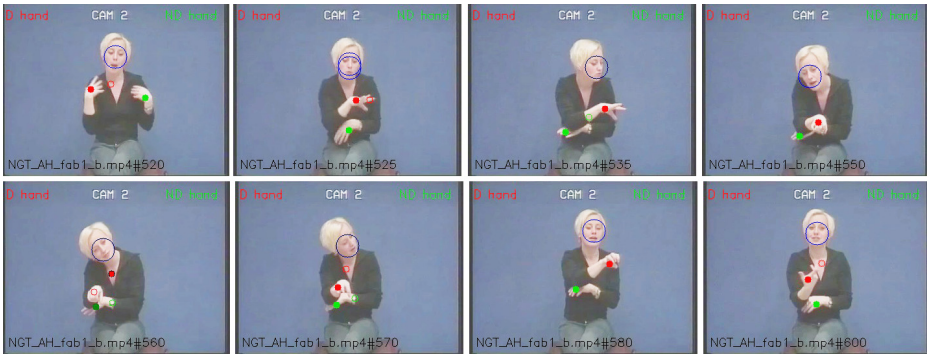

Fig. 3. Tracking in the presence of complex hand interactions and cross-over events.

exhibits a marginal increase in the number of FPs over regular MHT. While the reason behind this could not be ascertained, our analysis of the tracking results showed that a large number of the FP observations are caused by a too simplistic representation of hand motion, mainly the affine motion model adopted for KLT feature clustering and the constant-velocity Kalman filters. Signing exhibits abrupt hand motion with many discontinuities in hand trajectories – something which cannot be easily modelled with affine motion and constant-velocity KFs. To minimise feature cluster fragmentation and the FPs that arise from it, we could potentially use an IMM (interacting multiple model) approach, where several KFs, tuned to different hand manoeuvres, are run in parallel. In [25], an IMM is applied for tracking the hands in a natural conversation context.

6 Conclusion

We have proposed a mechanism for integrating *prior* knowledge about hand motion constraints described by sign language phonological models into our MHT-based tracking framework, in order to provide better tracking robustness especially in the presence of occlusion and hand interaction events. The constraints are integrated within the hypothesis evaluation mechanism of MHT and defined in terms of probabilistic density maps. We demonstrated this approach via the implementation of one such constraint – the hand symmetry condition. Experimental results demonstrate the effectiveness and the prospect of our approach, especially in improving tracking accuracy. And since hand tracking is central to sign recognition, it is expected that our approach will show a marked improvement in sign recognition once it is incorporated within the ASLR process.

Future work will look into: (1) adding more constraints into the tracking process from sign language phonological models (e.g. “h2-P”); (2) look into the use of handshape information both in the tracking process and the addition of phonological constraints related to the handshape; (3) integrate the hand motion constraints in a more principled and structured way, perhaps adopting

a parametric or semi-parameteric probabilistic approach instead of the current non-parametric representation; (4) and employ better motion models for tracking the hands, instead of the current use of affine motion models and constant velocity Kalman filters, perhaps using IMMs to handle abrupt hand motions.

References

1. von Agris, et al.: Recent developments in visual sign language recognition. Universal Access in the Information Society **6**(4), 323–362 (2007)
2. Sandler, W., Lillo-Martin, D.: Sign Language and Linguistic Universals. Cambridge Univ. Press (2006)
3. Bar-shalom, Y., Daum, F., Huang, J.: The Probabilistic Data Association Filter. IEEE Control Syst. Mag., 82–100 (2009)
4. Blackman, S.: Multiple Hypothesis Tracking For Multiple Target Tracking. IEEE Aerosp. Electron. Syst. Mag. **19**(1), 5–18 (2004)
5. Roshtkhar, M., et al.: Multiple Object Tracking Using Local Motion Patterns. BMVC (2014)
6. Amer, A.: Voting-based simultaneous tracking of multiple video objects. IEEE Trans. Circuits Syst. Video Technol. **15**(11), 1448–1462 (2005)
7. Shafique, K., Shah, M.: A Non-Iterative Greedy Algorithm for Multi-frame Point Correspondence. IEEE Proc. Int. Conf. Comp. Vision, 110–115 (2003)
8. Klinger, A., et al.: A Dynamic Bayes Network for Visual Pedestrian Tracking. ISPRS **40**(3), 145–150 (2014)
9. Reid, D.: An algorithm for tracking multiple targets. IEEE Trans. Automat. Contr. **24**(6), 843–854 (1979)
10. Yilmaz, A.: Object Tracking: A Survey. ACM Comput. Surv. **38**(4), 1–45 (2006)
11. Ragland, K., Tharcis, P.: A Survey on Object Detection, Classification and Tracking Methods. IJERT **3**(11), 622–628 (2014)
12. Thirde, D., et al.: Robust Real-Time Tracking for Visual Surveillance. EURASIP J. Advances in Signal Proc **2007**(1), 1–23 (2007)
13. Ying, L., Xu, C., Guo, W.: Extended MHT algorithm for multiple object tracking. In: Proc. Int. Conf. Internet Multimedia Computing and Service (2012)
14. Lao, Y., Zheng, Y.: Tracking highly correlated targets through statistical multiplexing. Image and Vision Computing **29**(12), 803–817 (2011)
15. Veenman, C., Reinders, M., Backer, E.: Resolving motion correspondence for densely moving points. Trans. PAMI **23**(1), 54–72 (2001)
16. Antunes, D., et al.: A Library for Implementing the Multiple Hypothesis Tracking Algorithm. CoRR **23**(1), 54–72 (2011)
17. Cox, I.: Miller, M: On Finding Ranked Assignments With Application to Multi-Target Tracking & Motion Correspondence. Trans. Aerosp. Electron. Syst. (1996)
18. Wimmer, M.: Adaptive skin color classificator. In: Proc. GVIP, pp. 324–327 (2005)
19. Tomasi, C., Shi, J.: Good Features to Track. In: Proc. CVPR, pp. 593–600 (1994)
20. Doermann, D: Tools and Techniques for Video Performances Evaluation (2000)
21. Crasborn, O., et al.: ECHO Data Set for Sign Language of the Netherlands (NGT) (2004)
22. Arras, K., et al.: Efficient People Tracking in Laser Range Data using a Multi-Hypothesis Leg-Tracker with Adaptive Occlusion Probabilities. Int. Conf. on Robotics and Automation (2008)

23. Kasturi, R., et al.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video. IEEE PAMI **31**(2), 319–336 (2009)
24. Amditis, A., et al.: Multiple Hypothesis Tracking Implementation, 199–220 (2012)
25. Wu, S., Hong, L.: Hand tracking in a natural conversational environment by the interacting multiple model and probabilistic data association (IMM-PDA) algorithm. Pattern Recognition **38**(11), 199–220 (2005)
26. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: Proc. Int. Conf. Image Processing (2002)