Multiple Hypothesis Tracking with Sign Language Hand Motion Constraints

Mark Borg, Kenneth P. Camilleri

Department of Systems & Control Engineering, Faculty of Engineering, University of Malta

UNIVERSITY OF MALTA
L-Università ta' Malta

Vision-based Automated Sign Language Recognition

Hand tracking is a critical component of ASLR systems, 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 due to the presence of noise and high variability in scene conditions

Visual Complexity

- *complex hand movements*
- *frequent hand occlusions*
- *signing occurs in a confined space*
- *fast motion blur*
- *detection errors (hand, skin, etc.)*

Multiple Hypothesis Tracking

MHT is a probabilistic framework that offers a high degree of robustness in the noisy and visually complex multi-object tracking setting of ASLR. During each tracking step, the MHT algorithm keeps multiple hypotheses about potential *hand – observation* matches (**data association**). These hypotheses are then propagated into the future in anticipation that subsequent observations will increase the probability of correct hypotheses.

Hypothesis Tree Explosion!

If unchecked, the MHT algorithm can suffer from an exponential growth of tracking hypotheses. Pruning of the hypothesis tree is thus required, such as employing a sliding temporal window, as well as discarding low-probability hypotheses. This also ensures the real-time operation of MHT. Managing in an effective manner the hypotheses and associated probabilities is a *key* factor to MHT's success.

Our Solution

Use of *prior* knowledge and hand motion constraints

managing Hypothesis Growth in MHT

Our solution incorporates prior knowledge about the constraints on hand motion as described by sign language linguistic models. These constraints are incorporated within the probabilistic framework of the MHT algorithm and are exploited by the MHT-based tracker in order to reduce the space of possible hypotheses.

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 in the presence of frequent and persistent hand occlusion events.

The hand motion constraints are incorporated within the hypothesis evaluation equation of the MHT algorithm via the use of probabilistic density maps.

Sign Language Phonological Models

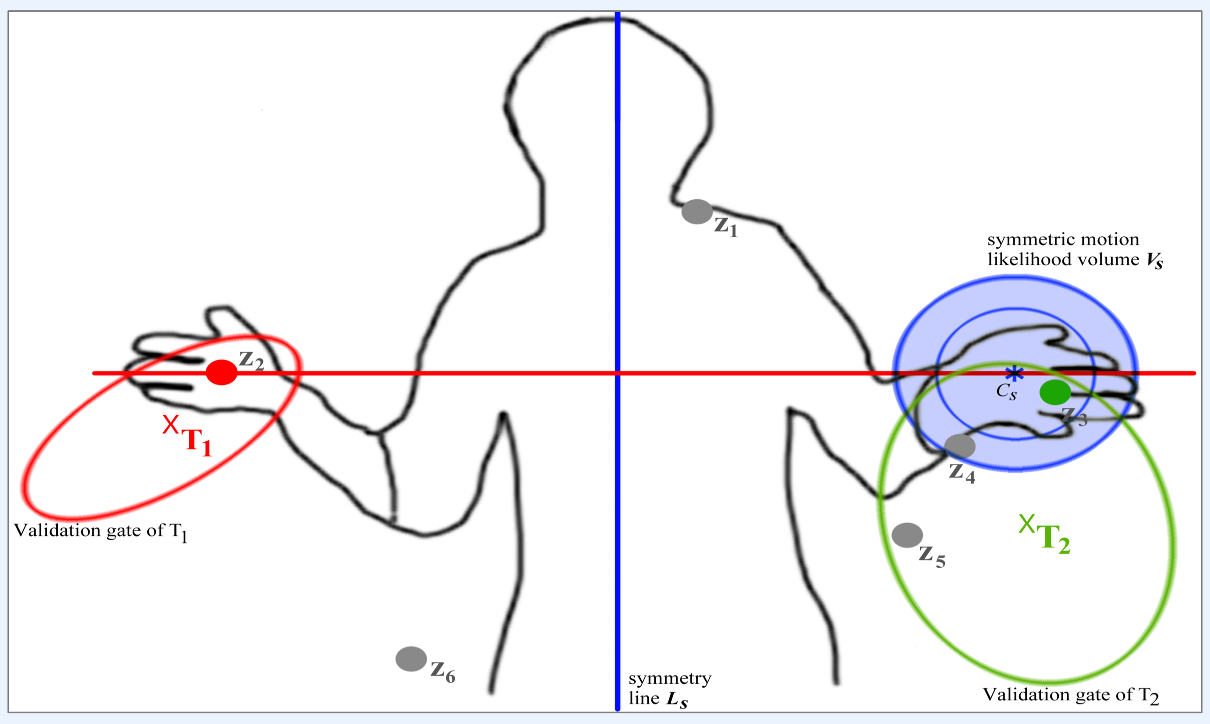
In Sign Language phonological models, the ***Hand Symmetry*** and ***Hand Dominance conditions*** limit the role of the non-dominant hand (*h2*) of the signer to serve as either a duplicate articulator (giving rise to the so-called *h2-S signs*), or as a place of articulation for the dominant hand (so-called *h2-P signs*).

Hand Symmetry constraint

In *h2-S signs*, the articulation of the non-dominant hand is symmetric to that of the dominant hand, and both must have the same handshape.

Hand Dominance condition

In *h2-P signs*, the non-dominant hand (*h2*) is stationary and the dominant hand (*h1*) moves while using *h2* as a place of reference against which the motion is performed.



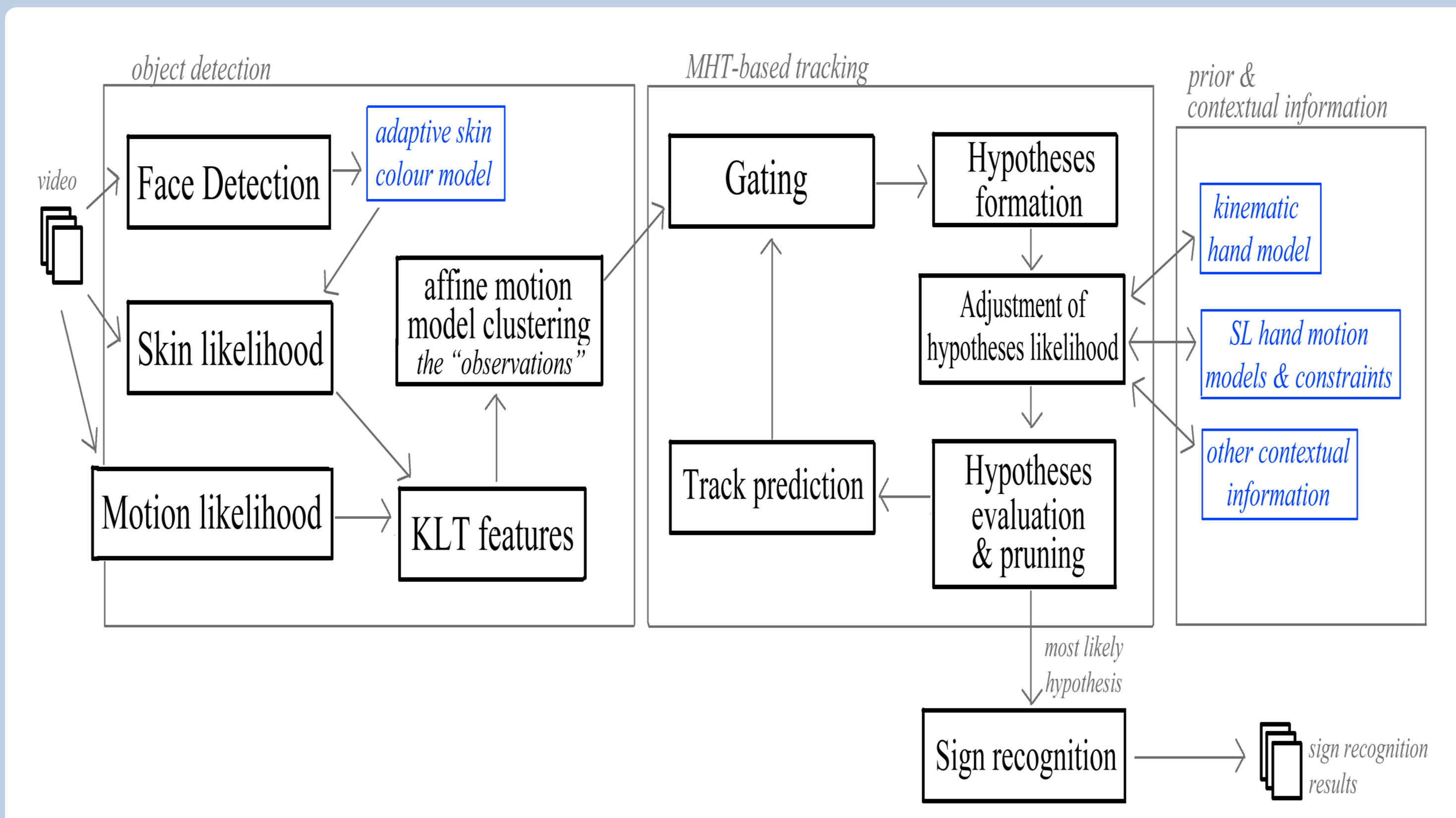
Implementation of our MHT-based system

Our system performs Face Detection both for face localisation purposes, as well as for performing system initialisation, such as that of learning the skin colour model of the signer. An Adaptive Skin Colour classifier is used to generate a Skin Likelihood Map. This is used in conjunction with a Motion Likelihood Map (generated via weighted frame differencing), to serve as an initial filtering stage for later feature extraction during the tracking stage.

The motion-based features used in our system consist of clusters of KLT features (corners) exhibiting a similar affine motion model. A multi-scale sparse optical flow algorithm is used for motion estimation of the features and the RANSAC scheme is used for robust affine motion model fitting. The clusters of KLT features and their associated affine motion models constitute the *observations* that are fed to the MHT stage of our tracking system.

We represent *targets* (the hands and face) by their affine motion model, centroid, and their spatial extent. A constant-velocity Kalman filter is then used for predicting hand motion, predicting hand occlusion via bounding box overlap, as well as smoothing the affine motion model for KLT feature clustering & replenishment.

During the MHT hypothesis generation stage, we employ the hand motion constraints in order to manage the growth of the hypothesis tree. Given a hypothesis ψ_k that associates the dominant hand target T_i with observation z_j, we employ the *Hand Symmetry Constraint* by locating a corresponding point for the non-dominant hand in the signing space reflected by the line of symmetry (L_s) anchored on the signer's body. This gives rise to the symmetric motion likelihood volume V_S, represented as a probability density map, which is then incorporated within the MHT's hypotheses evaluation & pruning mechanism as a weighted combination of the standard MHT's observation-target probability and the probability given by our density map.



Results

Our system was tested on the videos from the ECHO Sign Language (NGT) Corpus and evaluated using the CLEAR metrics MOTP and MOTA.

MOTP (multiple object tracking precision) measures how well the positions of the hands are estimated by the tracker, while MOTA (multiple object tracking accuracy) measures the number of mistakes that the tracker makes in terms of missed objects, false positives, and the number of label/identity switches that occur – an indication of the tracker's performance at keeping accurate trajectories in the presence of noise and occlusions.

Method	MOTP	MOTA
MHT	0.603	0.281
MHT + hand motion constraints	0.601	0.313

Our results show that our proposed approach of integrating sign language–based hand motion constraints within MHT, provides 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.

Conclusion & Future Work

We have successfully demonstrated a mechanism for integrating knowledge about hand motion constraints, based on sign language phonological models, into our MHT tracking framework, thus providing better tracking robustness especially in the presence of frequent occlusions and complex hand interaction events.

Future work will consider including more context information and hand motion constraints, tackle the issue of activating and switching between different constraints based on the signing discourse, as well as investigate the automatic learning of such constraints.