Context-Awareness and Selective Attention for Persistent Visual Tracking

Ying Wu
Electrical Engineering & Computer Science
Northwestern University
Evanston, IL 60208

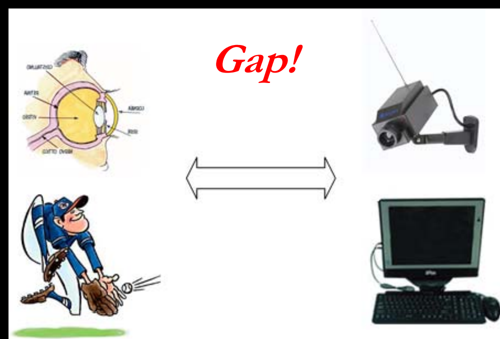
http://vision.eecs.northwestern.edu
yingwu@northwestern.edu

"Hopeless" for tracking?

2

A Huge Gap

Incredibly *easy* *vs.* Surprisingly *hard*



- Human visual perception seems to be attentional and selective.
- But most computational models for visual tracking appear to be over-simplified, and thus confronted.
- How can we bridge the gap?

3

Visual Attention

- *Visual attention* : cognitive processes to recruit resources for processing selected aspects of the retinal image more fully than non-selected aspects.
- *Spatial selection :*
  - one important aspect of visual attention.
  - the selectivity that samples the retinal image and processes a restricted region at eye fixation .
  - the so-called “mind’s eye”.
- Can this be modeled computationally?

4

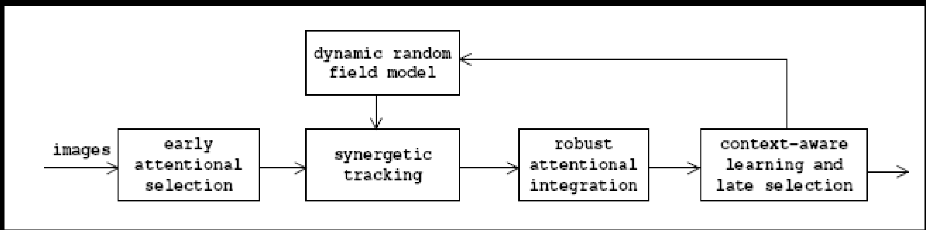
Spatial Attentional Selection

- *Early selection*
  - *Innate* principles
  - Performing initial pre-filtering in the very early stage.
  - *e.g.* attend to the moving objects.
- *Late selection*
  - Principles *learned* via experiences
  - Involving higher level processing.
  - *e.g.* learn the differences among camouflage objects.

5

Synergetic Selective Attention Model

- Synergetic selective attention (SSA) model
  - Early attentional selection
  - Synergetic tracking
  - Robust integration
  - Context-aware learning and late selection



6

Two case studies

- Selective attention
  - A general purpose tracker
- Context awareness
  - A powerful head tracker

Selective Attention

- A general purpose tracker
- Four components
  - Early selection
  - Synergetic tacking
  - Estimation integration
  - Context-aware late selection

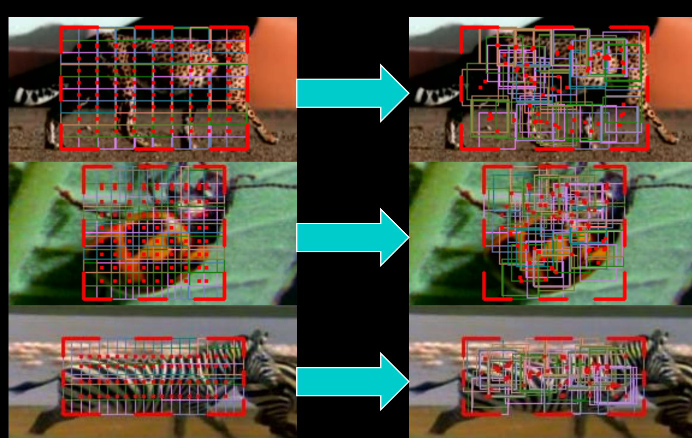
A general-purpose tracker

- Challenges:
  - no priors for the target
  - no off-line learning is available
  - unpredictable scenes and targets
    - ✓ Appearance/shape changes
    - ✓ camouflage distraction
    - ✓ complex partial occlusion
    - ✓ targets with irregular shapes

9

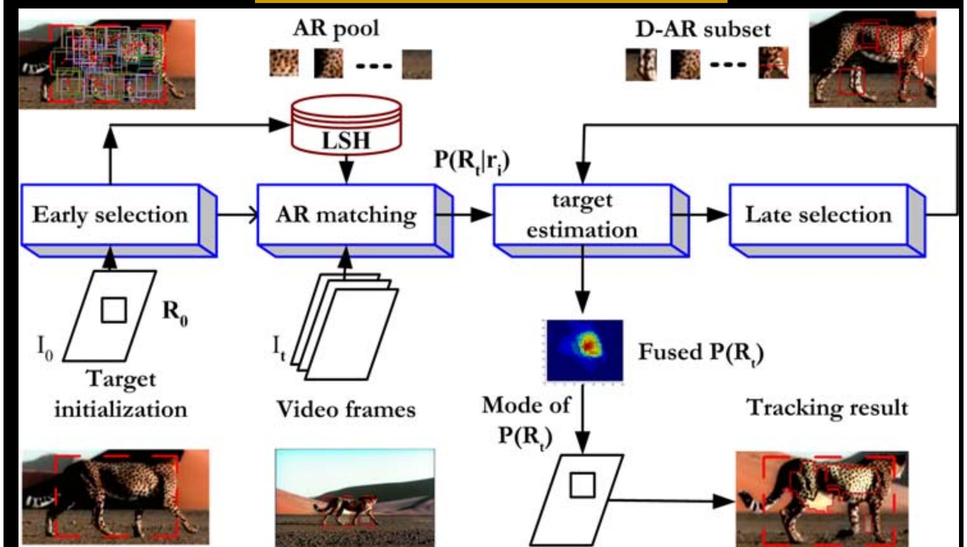
Attentional Regions

- Target representation: a pool of *attentional regions* (ARs) which are defined as salient image regions, *e.g.* those that have good localization properties.



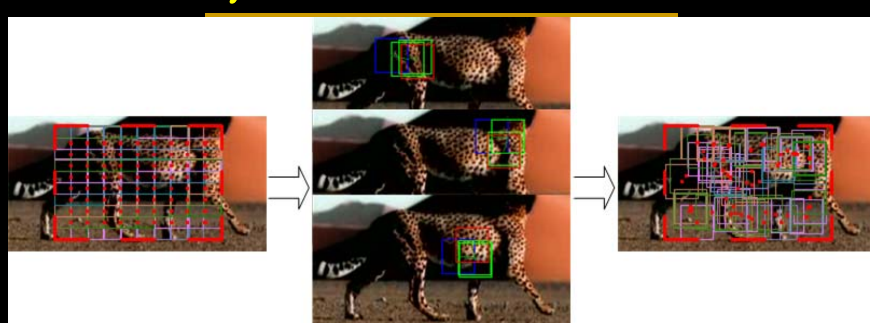
10

Attentional Visual Tracking



11

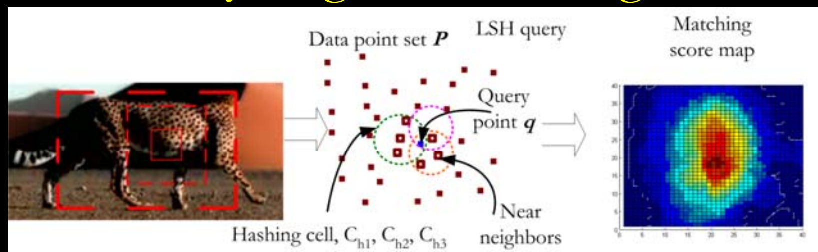
Early Attentional Selection



- Select ARs to be those that are sensitive to motion
  - Measuring the sensitivity (to motion estimation)
  - It is related to the condition number of a linear system (Fan & Wu: CVPR06)
  - Locate with an efficient gradient-based method

12

Synergetic Tracking



- Local exhaustive match for all ARs
  - Matusita metric for two histograms **x** and **y**:

$$d(\mathbf{x},\mathbf{y}) = \sum_{j}^{D} \|\sqrt{x_j} - \sqrt{y_j}\|^2$$

  - Locality-sensitive hashing (LSH) accelerates approximate nearest neighbor searching.
    - ✓Search complexity: sub-linear
    - ✓Overhead: pre-indexing

13

Two Options

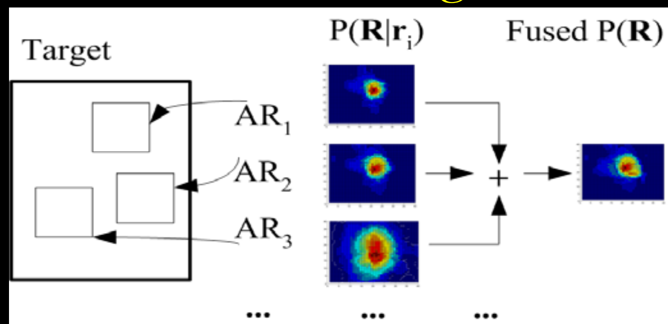
- LSH in tracking : both *indexing* and *query* costs.

√ Option A: L<N   Option B: L>N

- Complexity of attentional region matching.
  - ✓# of ARs: N (N<100)
  - ✓# of candidate regions: M (M<3000)
  - ✓Exhaustive search: O(MN)
  - ✓# of hashing functions in LSH: L (10-20)
  - ✓LSH indexing and query: O(ML+NL)
  - ✓*$r \approx (L/N+L/M) \approx L/N$*
  - ✓E.g. N=100, L=20 r=0.2, N=36, L=10, r=0.28.

14

Estimation Integration

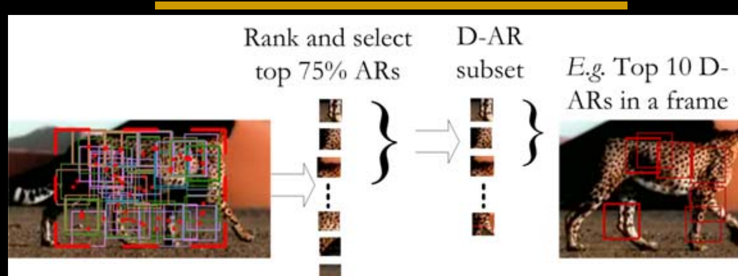


- P(**R**$_t$): the distribution of target location.
- P(**R**$_t$| **r**$_i$) : the conditional distribution of target's location given each AR **r**$_i$
  - Approximated by AR matching through LSH.
- Fuse P(**R**$_t$| **r**$_i$) by

$$\hat{P}(\mathbf{R}_t) \approx \sum_{i}^{\hat{N}} P(\mathbf{R}_t|\mathbf{r}_i)P(\mathbf{r}_i)$$

15

Context-aware Late Selection



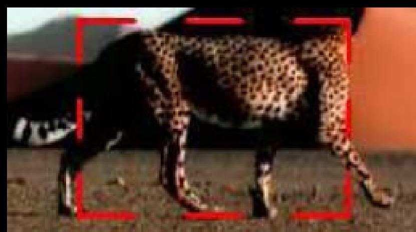
- Select more discriminative ARs (D-ARs) on-the-fly.
  - Measure the discrimination of each AR.
  - Rank the cross-entropy in a temporal sliding window.

$$KL(P(\mathbf{R}_t|\mathbf{r}_i)||\hat{P}(\mathbf{R}_t)) - KL(P(\mathbf{R}_t|\mathbf{r}_i)||P(B))$$

$$\tilde{H}(\mathbf{r}_i, \mathbf{R}_t) = \sum_{j=0}^{\Delta t} \beta^j H(P(\mathbf{R}_{t-j}|\mathbf{r}_i), \hat{P}(\mathbf{R}_{t-j}))$$

16

Two Selections



Target initialization

Early selection of ARs

Late selection of D-ARs during tracking

17

Discussion

- Early selection of an AR pool. ➡
  - Robust to small variations
    - ✓ lighting changes
    - ✓ small deformation.
- Late selection of a subset of D-ARs. ➡
  - Robust to
    - ✓ complex partial occlusion
    - ✓ inaccurate initialization due to irregular shapes.
- Local exhaustive search accelerated by LSH. ➡
  - Robust to quick motion

18

Experiment Settings

- Each AR is represented by
  - a color histogram in YCbCr space
  - 1040 bins with 32*32 for CbCr and 16 for Y.
- Acceleration by the integral histogram technique.
- 10-15 fps on average with C++ implementation tested on a PIV 3.0Ghz desktop.
- Real-world test sequences from *Google Video*.
  - People in crowd, walking, running, riding
  - Animals, *e.g.* cheetahs, zebras, and bug
  - Other targets, *e.g.* faces, ships, and bicycles

19

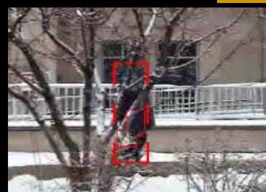
Examples

Pixels covered by more than one D-AR are highlighted.

20

Exciting Results

click to play video








- NYC street bicyclist
- occluding faces
- military vehicles

21

Context Awareness

- A powerful head tracker
- Four components
  - Early selection
  - Synergetic tacking
  - Estimation integration
  - Context-aware late selection

A "Hopeless" World

- Learn a target model for tracking???
- Enormous variation in the "hopeless" world
- Dilemma: efficient tracking v.s. effective verification
- How do you know the tracker is working?

23

The Light

- It is indeed hopeless
  - **If** the tracker only "looks" at the target per se
  - Clutter/occlusion
- See the light?
  - The target is not isolated
  - It is in context
  - Can its context help
  - Let's look back the example of head tracking …

24

Context-awareness Tracking

- Visual context
  - Early selection
  - what context is helpful?
- No prior of the context is available
- Discovering visual context on-the-fly
  - A learning/late selection process
- Tracking with visual context

25

Context: Auxiliary objects

- Three criteria for auxiliary objects:
  - Frequent co-occurrence with the target
  - Consistent motion correlation with the target
  - Suitable for tracking
- Note: auxiliary objects can be
  - solid semantic objects or image regions
  - close to the target or not
  - have intrinsic relations with the target or merely temporary correlations in a short period.
- To make things simple, we use rough color segments as auxiliary objects in this work.

26

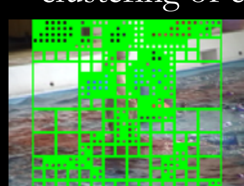
Sample auxiliary objects

Suppose the target is head, the red dash boxes indicate the auxiliary objects discovered automatically by data mining.

27

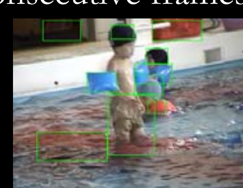
A specific implementation

- Rough color segments (color histogram and motion parameters}
- Split-and-merge quad-tree color segmentation
- Computationally efficient: 7-8ms for 256×256 image
- Heuristics that prune large or tiny color segments
- Mean-shift matching that facilitates incremental clustering of color segments in consecutive frames



Split stage

Merge stage



Color segments/item candidates

28

Four components

- Early selection
  - Identifying a set of color segments
- Late selection
  - Mining visual context
  - Learning auxiliary objects
  - Forming a dynamic random field
- Synergetic tracking
  - Inference on the RF through BP
- Robust integration/fusion
  - Removing outliers before fusion for verification

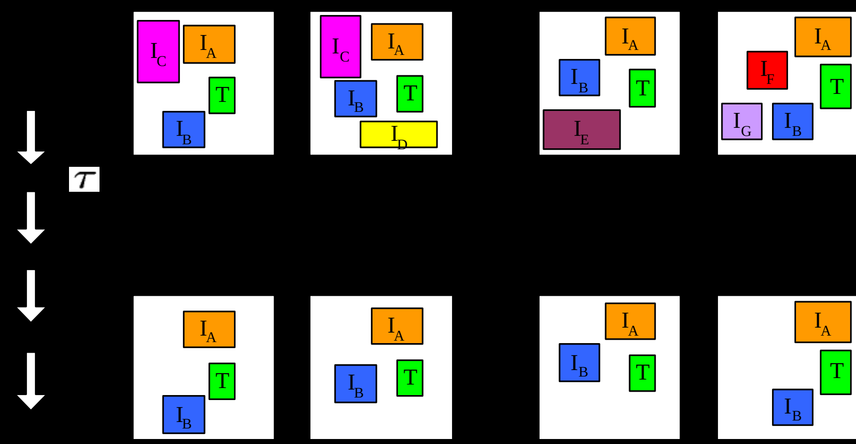
29

Mining Visual Context

- Item candidate generation
  - Extract simple image features as item candidates, e.g. rough color segments.
- Transaction generation
  - Quantize the item candidates by incremental clustering
  - Generate the transactions, i.e. matching the color segments in consecutive frames.
- Frequent item mining (FIM)
  - Frequent co-occurrence with the target
- Multibody grouping
  - Identify pair-wise motion correlation between the target and auxiliary objects

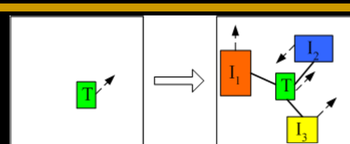
30

Illustration: Context Mining



31

Multibody Grouping



- Assume an affine motion model *A* between target *y* and candidate auxiliary object *x*: $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \mathbf{b}.$
- Subtract the mean $\tilde{\mathbf{y}}_t = \mathbf{y} - \bar{\mathbf{y}}$, $\tilde{\mathbf{x}}_t = \mathbf{x} - \bar{\mathbf{x}}$ and stack *y* and *x*, the covariance matrix can be expressed as:

$$\mathbf{C} = E[\begin{pmatrix} \tilde{\mathbf{y}}_t \\ \tilde{\mathbf{x}}_t \end{pmatrix} (\tilde{\mathbf{y}}_t^T, \tilde{\mathbf{x}}_t^T)].$$

- Rank of *C* indicates whether the two motions are correlated.

32

Multibody Grouping

Denote the covariance matrix of x as C^x, we have

$$\hat{\mathbf{C}} = \sum_{i=0}^{N} \begin{pmatrix} \tilde{\mathbf{y}}_{t-i} \\ \tilde{\mathbf{x}}_{t-i} \end{pmatrix} (\tilde{\mathbf{y}}_{t-i}^T, \tilde{\mathbf{x}}_{t-i}^T) = \begin{pmatrix} \mathbf{A}\hat{\mathbf{C}}^x\mathbf{A}^T + \sigma^2 & \mathbf{A}\hat{\mathbf{C}}^x \\ \hat{\mathbf{C}}^x\mathbf{A}^T & \hat{\mathbf{C}}^x \end{pmatrix}$$

Perform eigenvalue decomposition: $\hat{\mathbf{C}} = \mathbf{Q}\Lambda\mathbf{Q}$

The number of eigenvalues λ that are larger than the noise variance indicates if the candidate is an AO.

$$\#\text{ of } \{\lambda_j^2 \gg \sigma^2\} \begin{cases} > 2 & \text{NOT AO} \\ <= 2 & \text{AO} \end{cases}$$

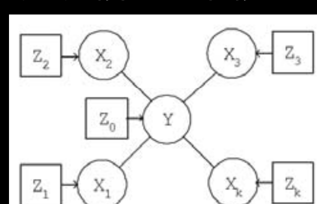
The affine motion can be solved by using the two least eigenvectors through subspace analysis:

$$\mathbf{A}^T \begin{pmatrix} q_{31} & q_{41} \\ q_{32} & q_{42} \end{pmatrix} + \begin{pmatrix} q_{33} & q_{43} \\ q_{34} & q_{44} \end{pmatrix} = 0.$$

33

Synergetic Tracking

- Tracking a random field



- Estimating based on belief propagation

$$\begin{aligned} p(\mathbf{y}|\mathbf{Z}) &\propto \hat{p}_0(\mathbf{y}|\mathbf{Z}) \prod_k m_{k0}(\mathbf{y}), \\ m_{k0}(\mathbf{y}) &= \int_{x_k} \hat{p}_k(\mathbf{x}_k|\mathbf{Z})\psi_{k0}(\mathbf{x}_k, \mathbf{y}) d\mathbf{x}_k, \end{aligned}$$

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{Z}) &\propto \hat{p}_k(\mathbf{x}_k|\mathbf{Z}) m_{0k}(\mathbf{x}_k) \quad k = 1, \ldots, K, \\ m_{0k}(\mathbf{x}_k) &= \int_y \hat{p}_0(\mathbf{y}|\mathbf{Z}) \prod_{\mathbf{x}_i \setminus \mathbf{x}_k} m_{i0}(\mathbf{y}) d\mathbf{y}, \end{aligned}$$

34

Robust fusion

- To identify the inconsistency of the trackers and remove the outliers before fusion are critical.
- The relative distances and scales between the target and auxiliary objects are modeled as Gaussians.
- Theorem: to detect pair-wise inconsistency between two Gaussian sources and if:

  where is the 2-norm conditional number of ,
  and they are consistent if :

  (Please refer to Gang Hua CVPR'06 paper)

35

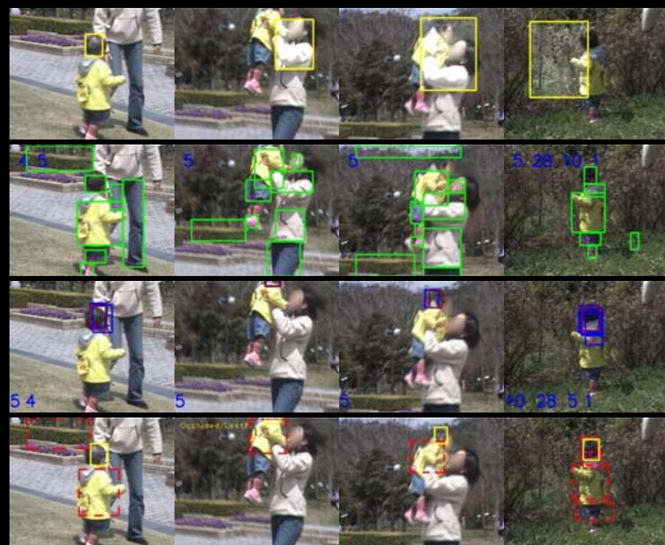
Experiment Settings

- Test data: amateur videos
- Target tracker: contour based head tracker.
- Auxiliary trackers: Mean-shift trackers in normalized R-G color space with 32×32 bins.
- Motion parameters: location and scales $\mathbf{x}=\{u,v,s_u,s_v\}$.
- C++ implementation: 5-10fps on Pentium IV 3G for 320×240 sequences.

36

An Example

- Single tracker
- Mining results
- Fusion results
- CAT tracker

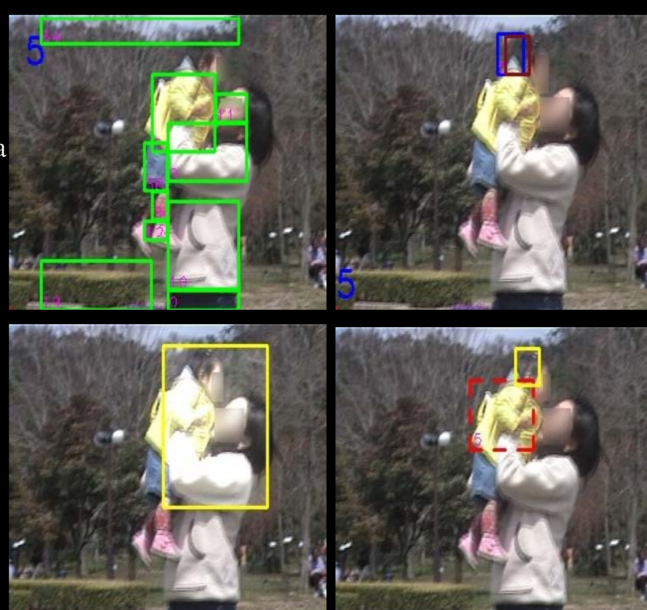


visual data mining

robust fusion

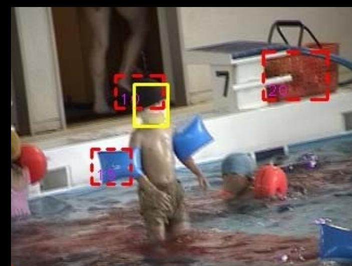
A dedicated head tracker (comparison)

CAT



Click to see video

More Promising Results

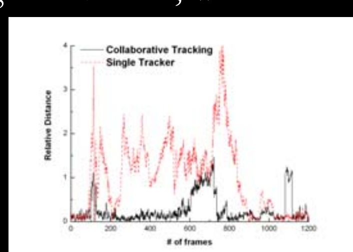


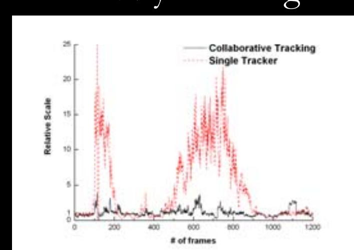
39

Quantitative evaluation

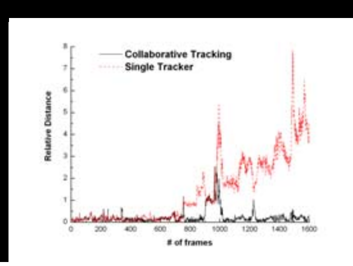
Quantitative evaluation: relative distances and scales to the ground truth, which are normalized by true target scales.

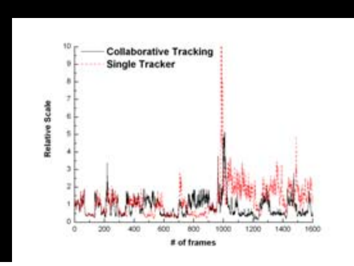
Kid in yellow





Dancing girl





40

Summary

- Selective attention and context awareness
- Synergetic selective attention model
  - Early selection
  - Synergetic tracking
  - Robust fusion
  - Context-aware late selection
- Future work
  - Context mining and learning (generative and discriminative)
  - The principle of early selection

41

Related Publications

- Ming Yang, Gang Hua and Ying Wu, "Context-Aware Visual Tracking", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.31, No.7, pp.1195-1209, July 2009
- Zhimin Fan, Ming Yang and Ying Wu, "Multiple Collaborative Kernel Tracking", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.29, No.7, pp.1268-1273, July 2007
- Ying Wu and Jialue Fan, "Contextual Flow", in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'09)*, Miami, FL, June 2009.
- Ming Yang, Junsong Yuan and Ying Wu, "Spatial Selection for Attentional Visual Tracking", *in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, MN, June 2007
- Ming Yang, Ying Wu and Shihong Lao, "Intelligent Collaborative Tracking by Mining Auxiliary Objects", *in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 06)*, New York City, NY, June 17-22, 2006.
- Jialue Fan, Jiang Xu and Ying Wu, "Context-aware Tracking of Small Targets in Video", *in Proc. Conf. on Signal and Data Processing of Small Targets, in SPIE Symposium on Optical Engineering and Applications*, San Diego, CA, August 2009

42