Automatic Visual Attention System
自動注視檢出

2009.11.23

朴 珉徹

Park, Min-Chul

韓國科學技術研究院

KIST SIAT

Contents

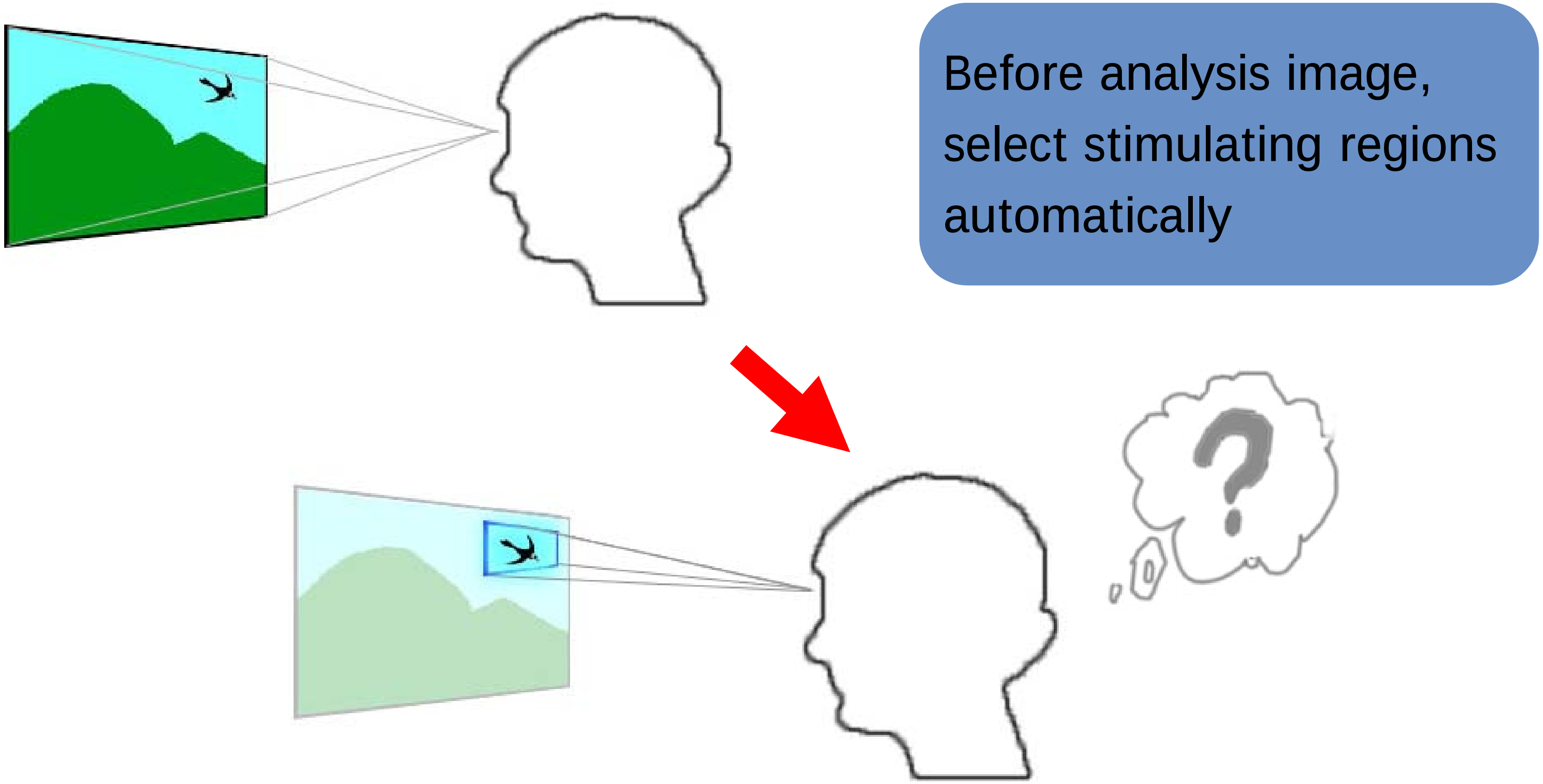
- Introduction
- Previous works
- Problems
- Object
- Proposed Methods
- Experiments
- Conclusion

Introduction

- **Perceptive Visual Attention Model**
  - Human behavioral system has some patterns when it makes a decision and human visual system also follows the characteristics
  - If we can provide users with some mechanism to support making a decision of the viewpoint it would be a very helpful interface
  - In the end it should be generated by a model that imitates human visual system to bring out some similar results we do

Introduction

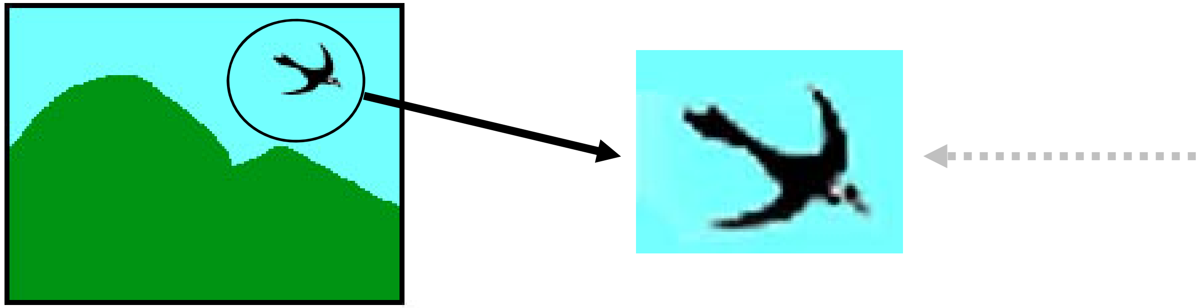
- **The human Perception**
  - Select attention region from a whole input image



Before analysis image, select stimulating regions automatically

Introduction

- **On Computer vision**



DataBase

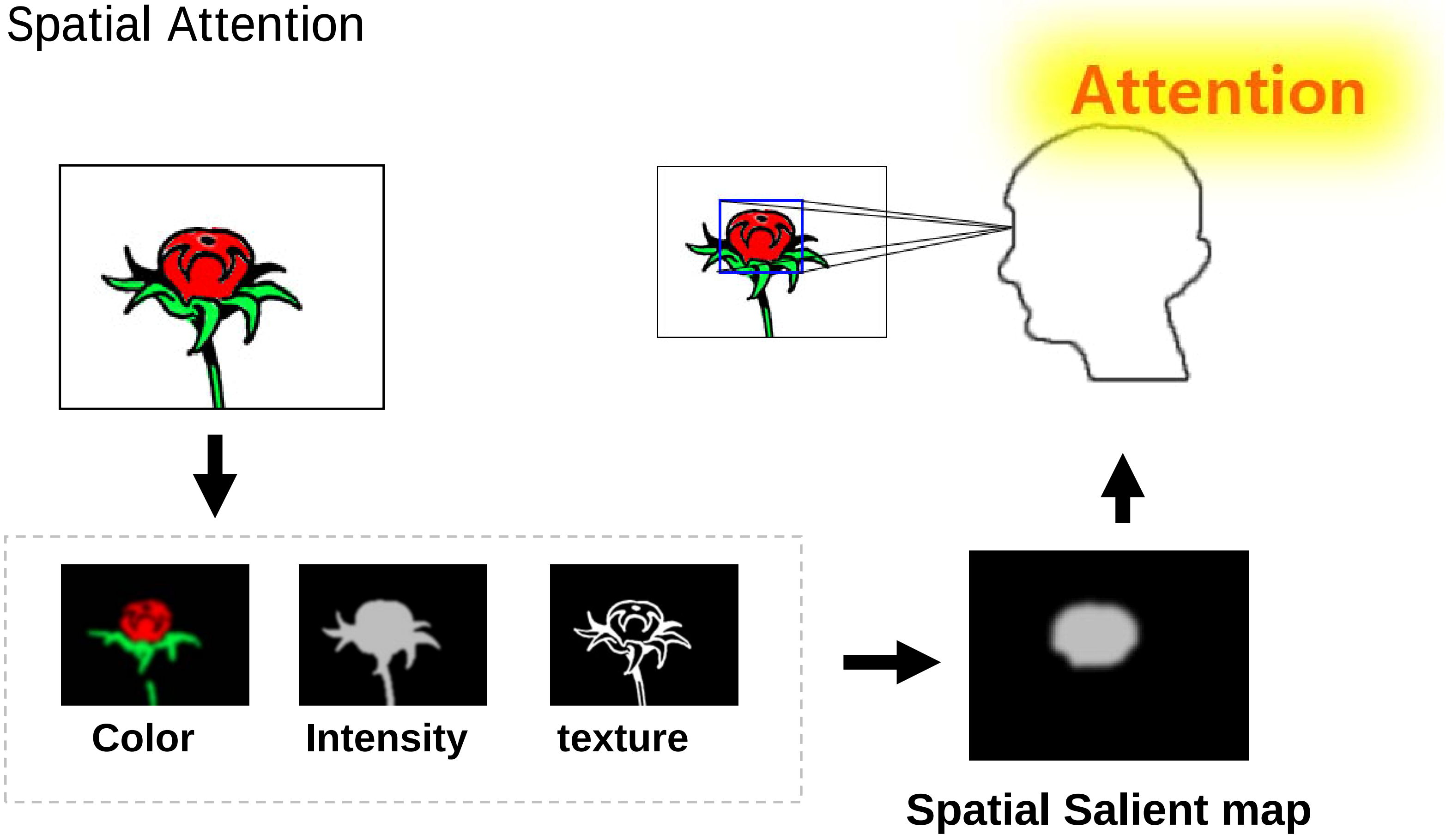
- Detects a certain region that attracts attention
  - The region is supposed of having certain useful information
    - Concentrates resource to some selected regions
    - Reduces computational burden, Uses resources efficiently

➡ Builds Visual Attention System similar to human perception

  - It becomes interesting topic in various research communities.
  - (3D Display, Multimedia processing, Computer Vision...)

Introduction

- **Using Spatial Saliency**
  - Spatial Attention



Attention

Color

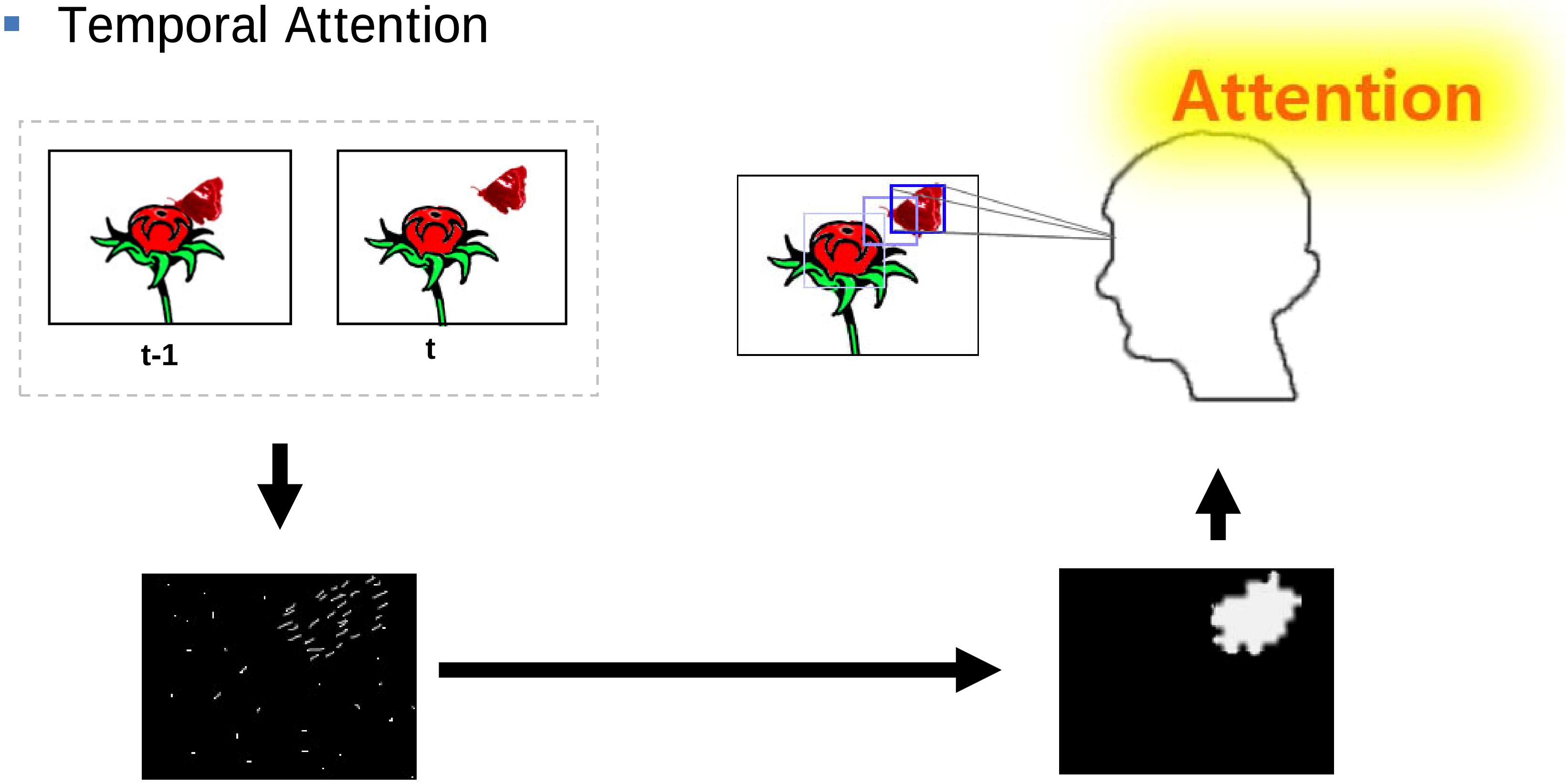
Intensity

texture

Spatial Salient map

Introduction

- **Using Temporal Saliency**
  - Temporal Attention



t-1

t

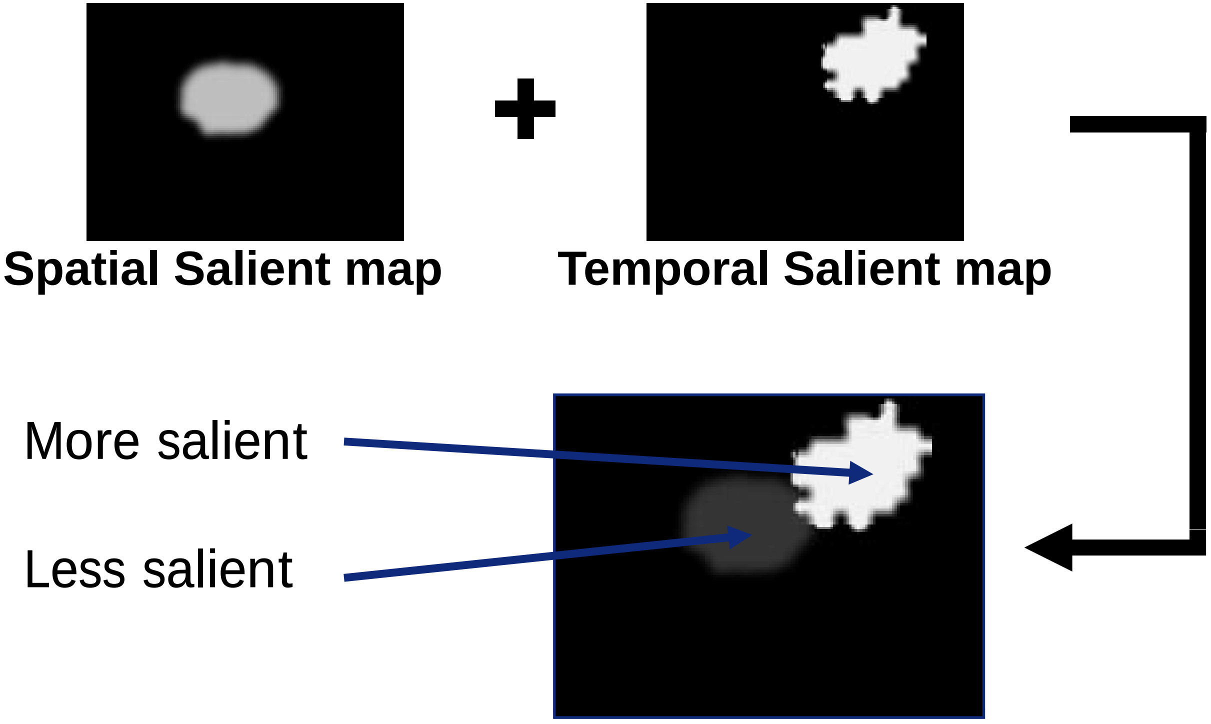
Attention

Motion vector

Temporal Salient map

Introduction

- **Using Both**
  - Integration of Spatial map and temporal map
  - Get Spatio-temporal Salient map



Spatial Salient map

Temporal Salient map

Various approaches...

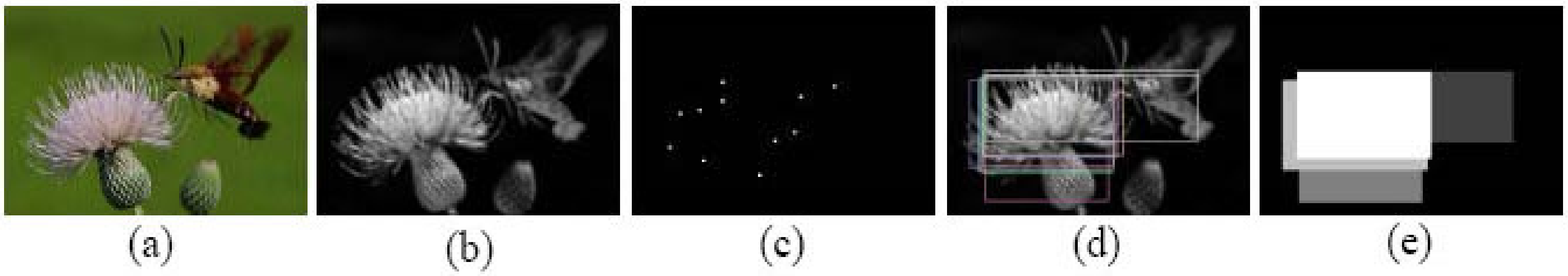
More salient

Less salient

Spatio-temporal Salient map

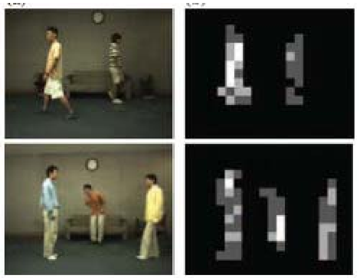
Previous works

- **Attention Detection in Video Sequences Using Spatiotemporal Cues (2006)**
  - Temporal attention
    - To detect motion, utilizes feature(interest) points instead of optical flow
    - Correspondences are established between the feature(interest)-points (frame(t-1), frame(t))
      - using Scale Invariant Feature Transformation (SIFT)
  - Spatial attention
    - Uses color histogram
    - Similar saliency value distribute widely -> overlapping attention points

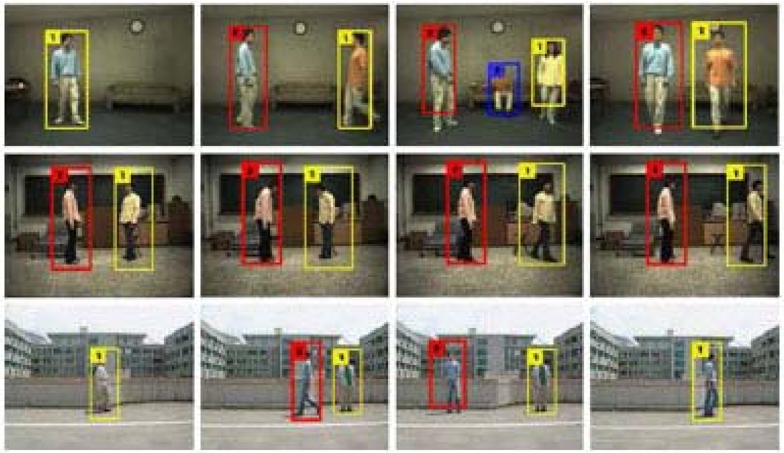


Previous works

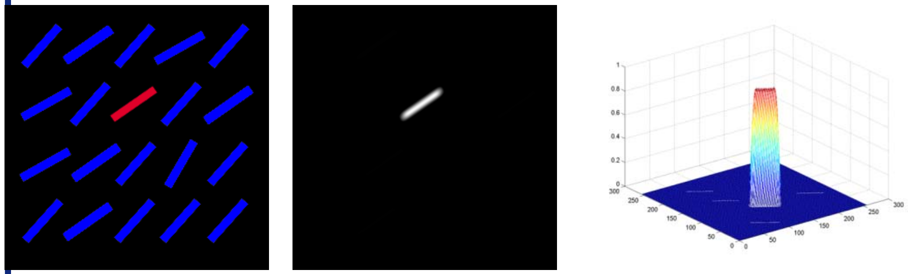
- **Salient Human Detection for Robot Vision (2007)**
  - Human detection based on Visual Attention System(VAS) using spatial saliency & temporal saliency
  - Temporal attention
    - Modified block matching method
      - Obtain probability distribution of motion
      - No consideration for direction of motion



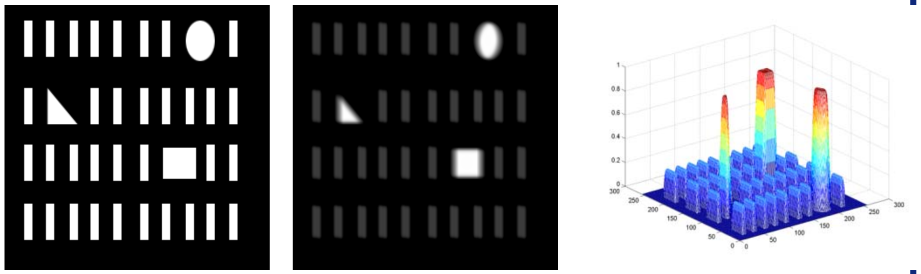
  - Example of VAS application for Robot vision



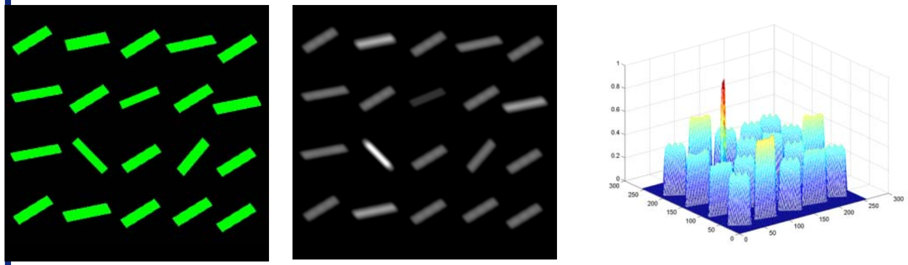
Our Previous Works



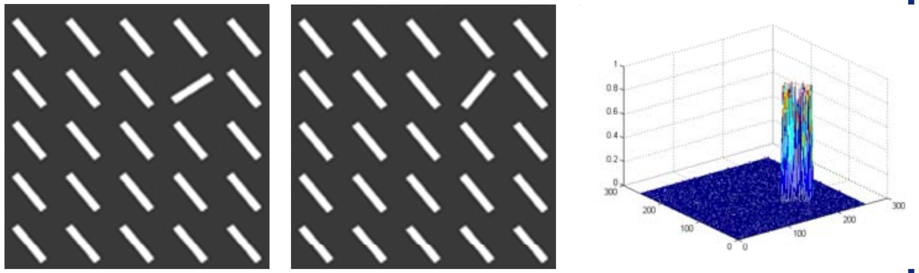
- Color



- Difference from surroundings

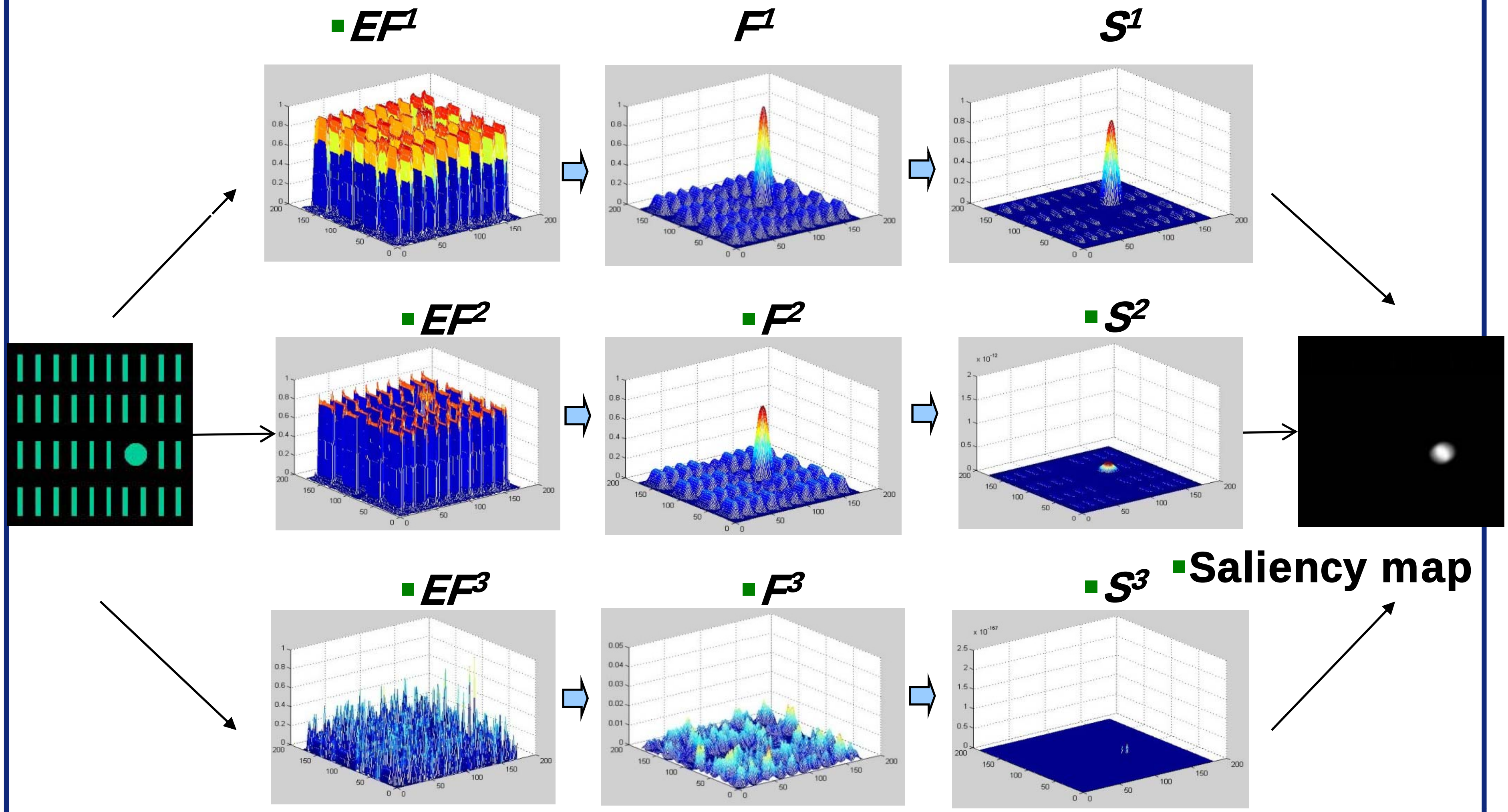


- Directions



- Motion

Feature Extraction and Integration Process



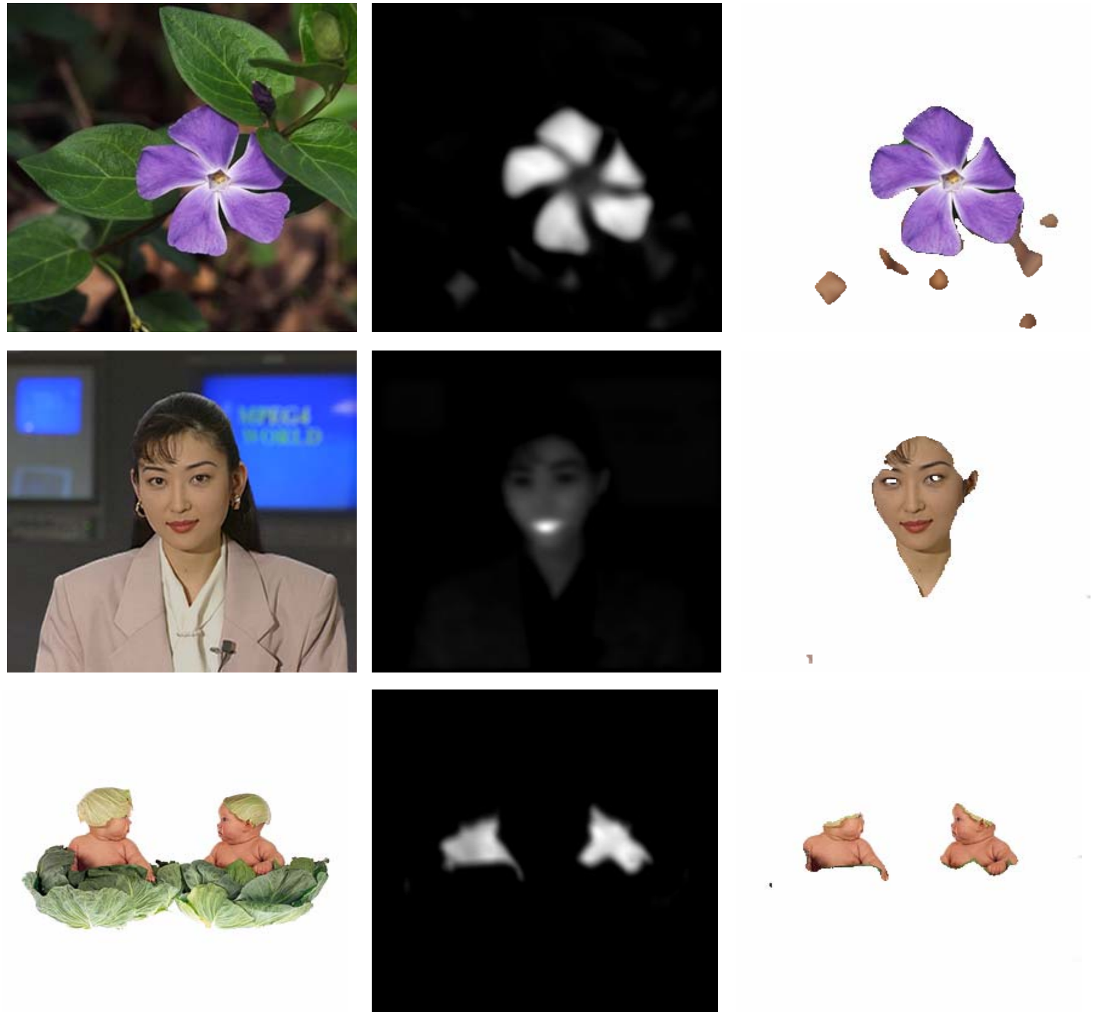
- EF^1

F^1

S^1

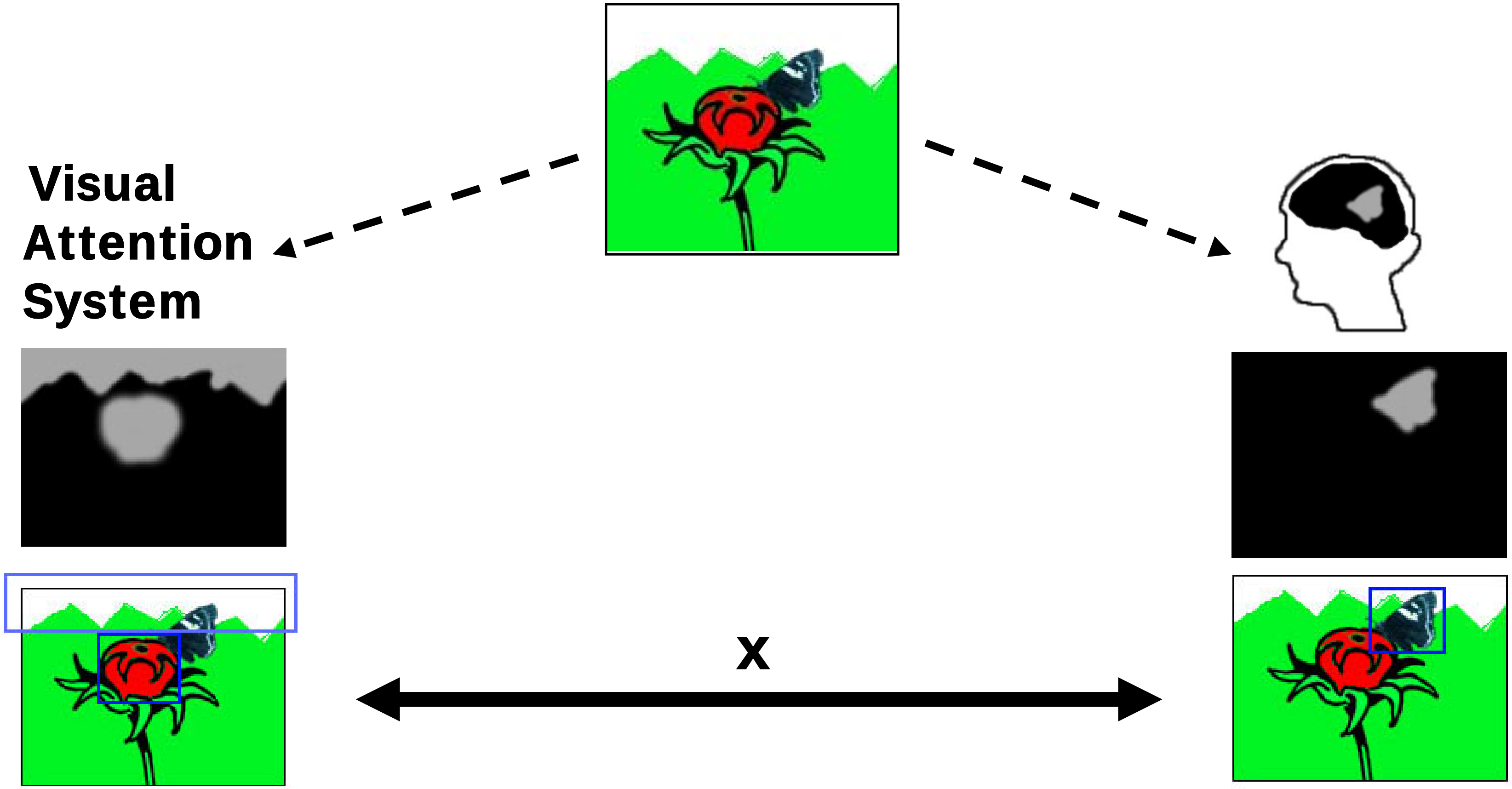
- EF^2
- F^2
- S^2
- EF^3
- F^3
- S^3
- Saliency map

Previous Works(Spatial Analysis)



Problems

- **Failure Case**



Problems

- **Problems**
  - Spatial salient map uses low-level features
    - Difficulty of understanding meaningful relations of objects
    - Different from the response of human perception
  - Compensate for spatial salient map by using motion information

➡ Reduction of noise information that deteriorates accuracy of temporal saliency

➡ How to define and make a difference among some salient regions under specific circumstance

- Some objects show similar features (motion, color…)
- Objects occlusion

Object

- **Prioritization and Segmentation**
  - Segmentation of salient regions
    - These regions may have similar spatial features, similar temporal features
    - Need to prioritize
  - When some objects are partially occluded or overlapped
    - Needs to keep individuality of each attention region
- **Effective Noise Elimination**
  - Eliminate meaningless(noise) motion information effectively
    - To obtain temporal salient map with meaningful motion information
    - Enhance Quality of Spatio-temporal salient map

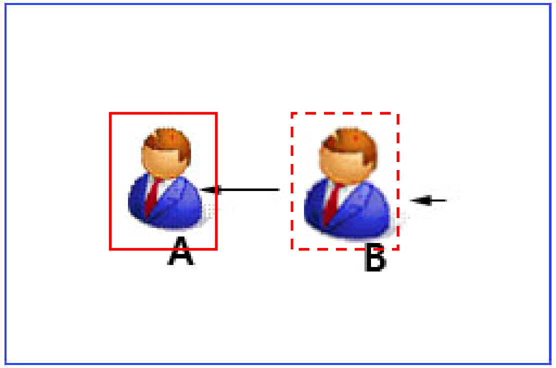
Proposed Method

- **Prioritization and Segmentation**
  - Spatial feature + Temporal feature + 3D-Depth information
    - 3D-Depth value : additional information of projected 2-D real world
    - To compensate 2-dimensional Spatiotemporal saliency map for improved results
  - Search minSAD (sum of absolute difference) block form stereo image
    - Generate disparity map
  - Segmentation Spatiotemporal salient map by shape of distribution of disparity value map
    - Disparity ↑ -> Priority ↑
    - Saliency value ↑ -> Priority↑

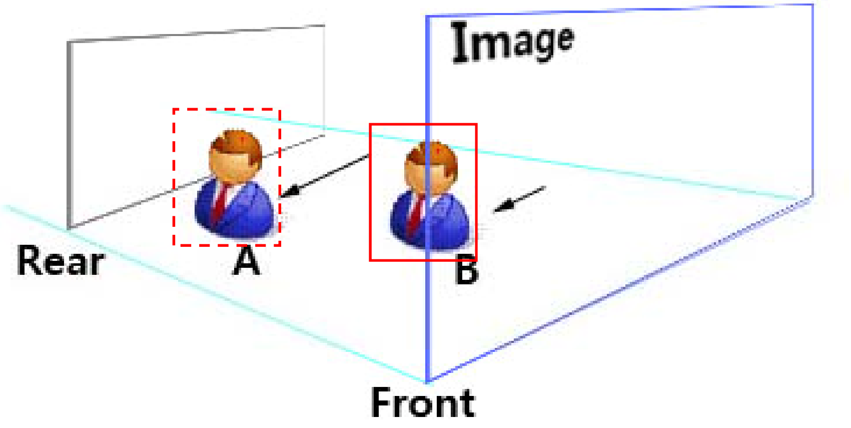
Proposed Method

- **Prioritization and Segmentation**

Without depth information



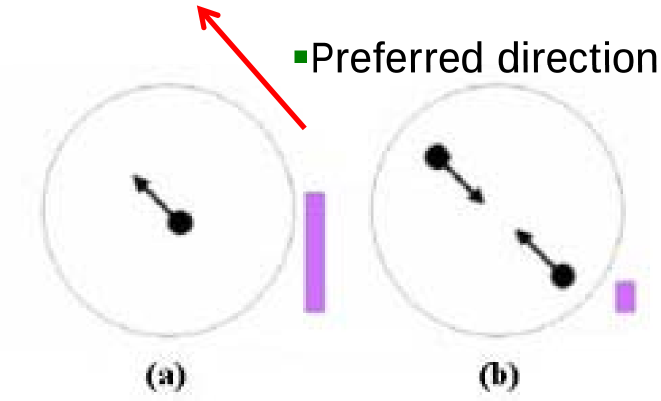
With depth information



Ex) Generally near one(object) to camera or viewpoints attracts stronger attention in real world

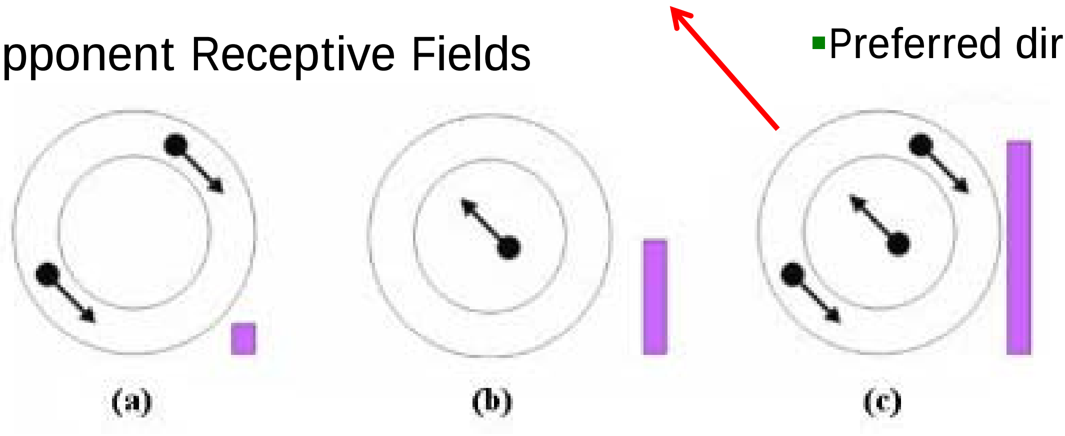
Proposed Method

- **Noise Elimination 1 (Based on psychological studies)**
  - Using properties of neurons in Middle Temporal cortex (MT)
    - Noise Filtration



- Preferred direction

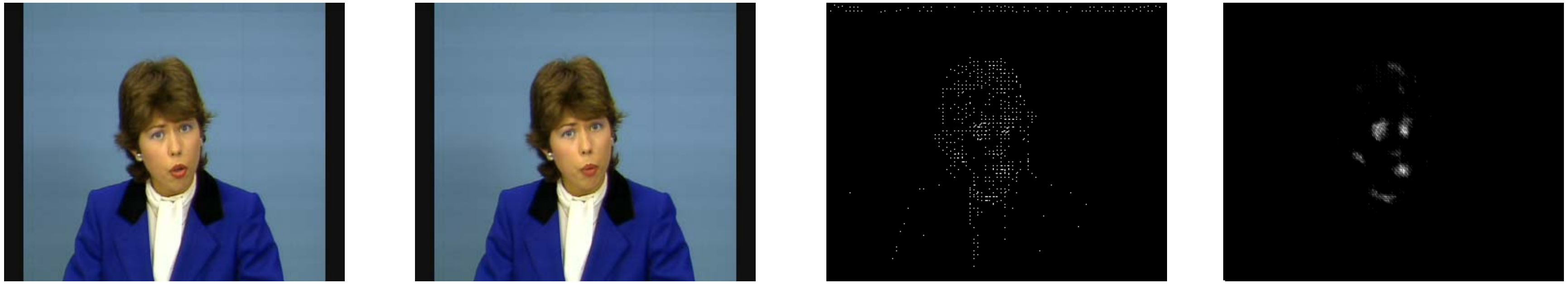
    - Double Opponent Receptive Fields



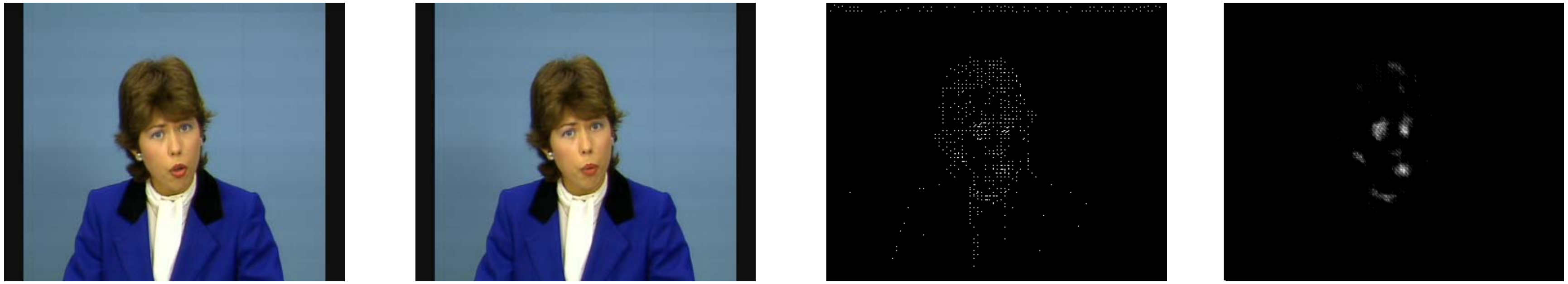
- Preferred direction

Experimental Results

In Claire image sequence, mouth and eye regions were marked as the most conspicuous regions (62.5%) from human experiments. If mouth and eye regions are included in the face region it takes up 70%.



Experimental Results



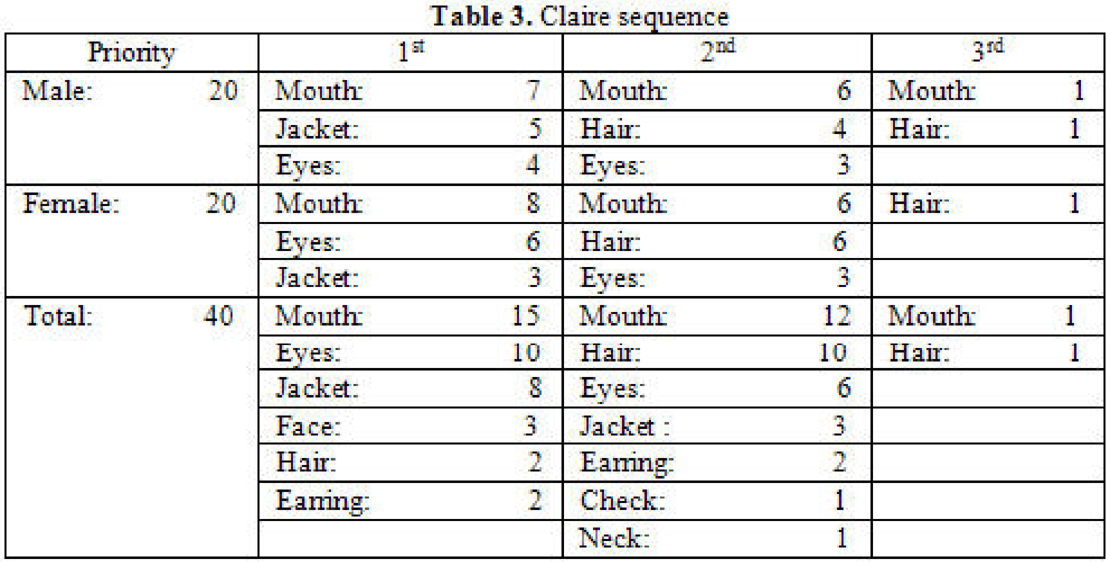
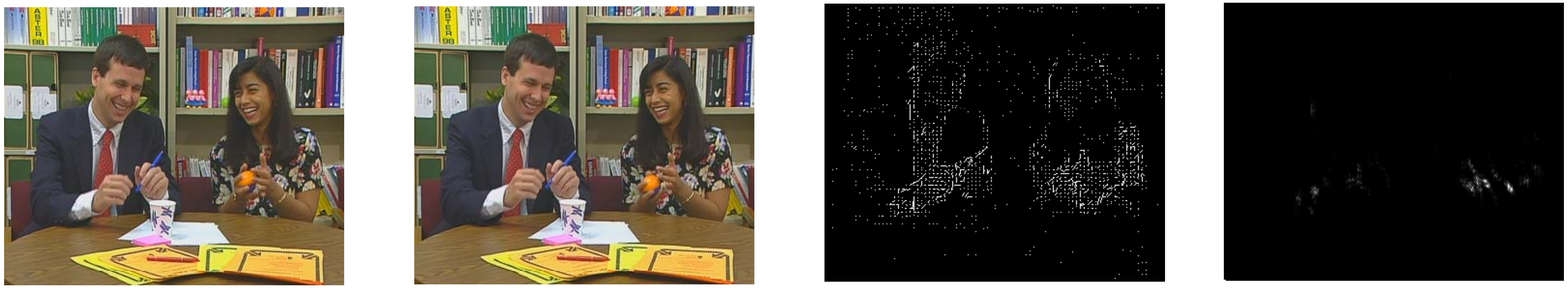


Table 3. Claire sequence

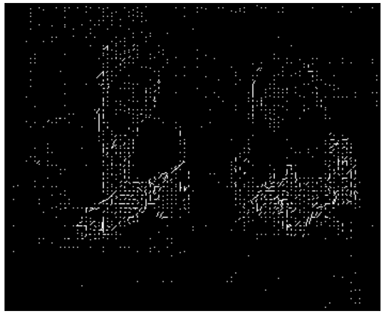
Priority		1st		2nd		3rd	
Male:	20	Mouth:	7	Mouth:	6	Mouth:	1
		Jacket:	5	Hair:	4	Hair:	1
		Eyes:	4	Eyes:	3		
Female:	20	Mouth:	8	Mouth:	6	Hair:	1
		Eyes:	6	Hair:	6		
		Jacket:	3	Eyes:	3		
Total:	40	Mouth:	15	Mouth:	12	Mouth:	1
		Eyes:	10	Hair:	10	Hair:	1
		Jacket:	8	Eyes:	6		
		Face:	3	Jacket :	3		
		Hair:	2	Earring:	2		
		Earring:	2	Check:	1		
				Neck:	1		

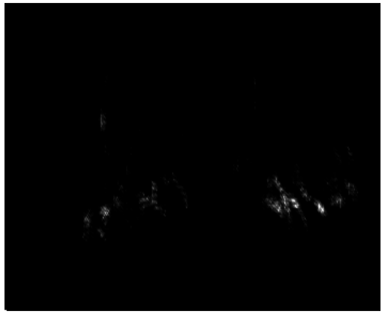
Experimental Results

In Pairs image sequence, ball region is marked as the most conspicuous one (60%) by human observers. The ball is detected with highest conspicuity also from the proposed attention module.



Experimental Results





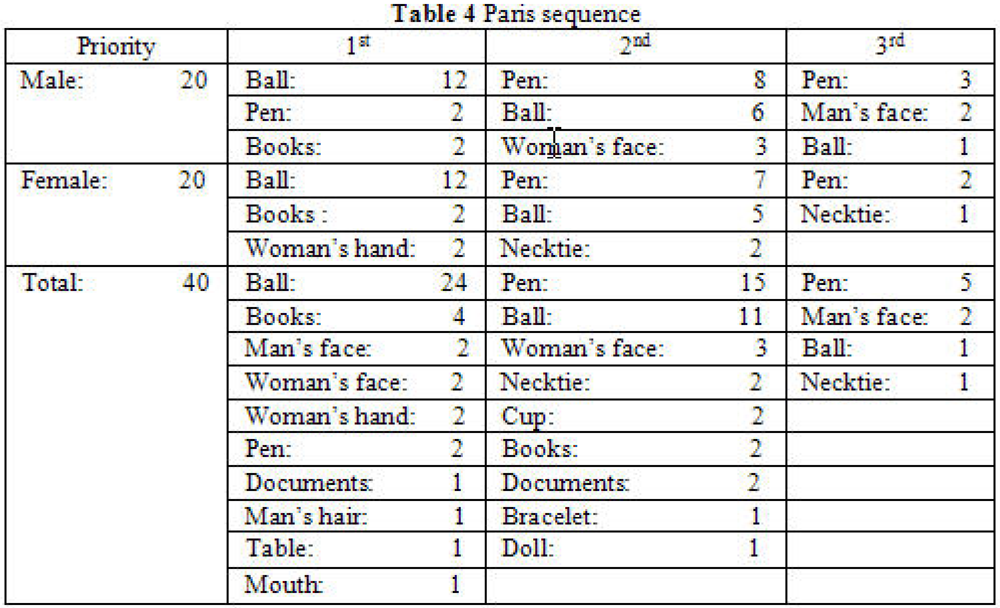
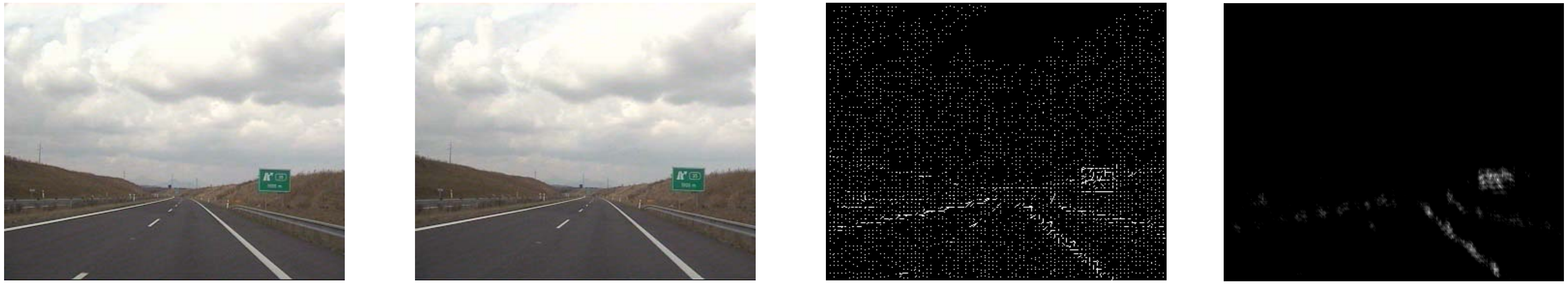


Table 4 Paris sequence

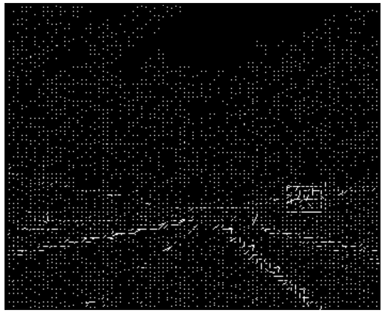
Priority		1st		2nd		3rd	
Male:	20	Ball:	12	Pen:	8	Pen:	3
		Pen:	2	Ball:	6	Man’s face:	2
		Books:	2	Woman’s face:	3	Ball:	1
Female:	20	Ball:	12	Pen:	7	Pen:	2
		Books :	2	Ball:	5	Necktie:	1
		Woman’s hand:	2	Necktie:	2		
Total:	40	Ball:	24	Pen:	15	Pen:	5
		Books:	4	Ball:	11	Man’s face:	2
		Man’s face:	2	Woman’s face:	3	Ball:	1
		Woman’s face:	2	Necktie:	2	Necktie:	1
		Woman’s hand:	2	Cup:	2		
		Pen:	2	Books:	2		
		Documents:	1	Documents:	2		
		Man’s hair:	1	Bracelet:	1		
		Table:	1	Doll:	1		
		Mouth:	1				

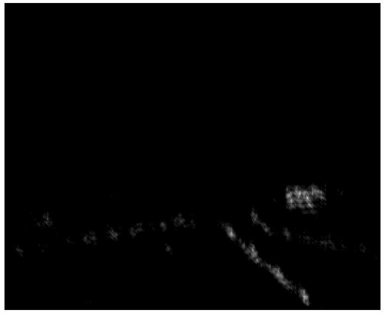
Experimental Results

In Highway image sequence, road and traffic signs are marked as the most conspicuous regions (87.5%). Same regions are also detected from the proposed attention module.



Experimental Results





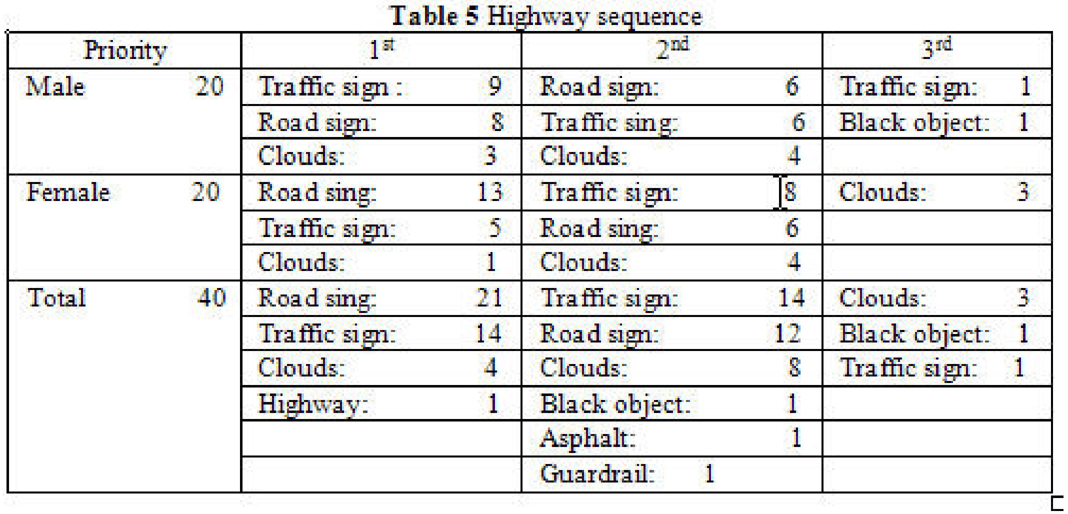
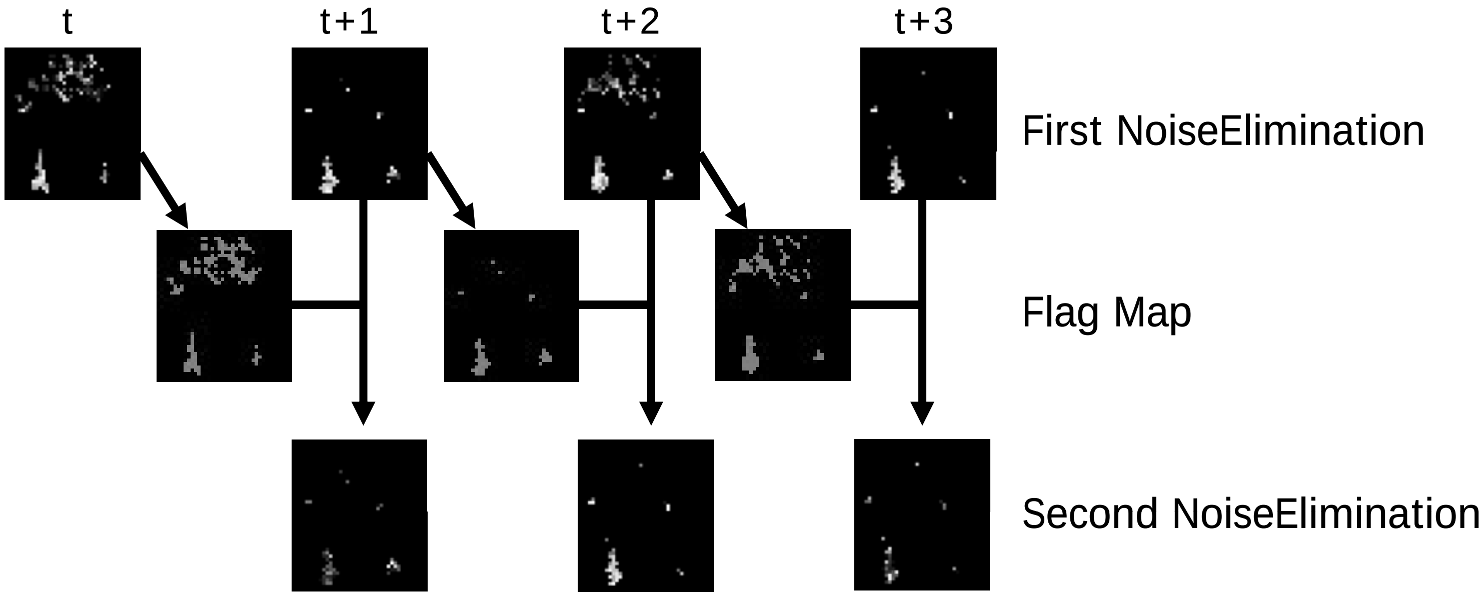


Table 5 Highway sequence

Priority		1^{st}		2^{nd}		3^{rd}	
Male	20	Traffic sign :	9	Road sign:	6	Traffic sign:	1
		Road sign:	8	Traffic sing:	6	Black object:	1
		Clouds:	3	Clouds:	4		
Female	20	Road sing:	13	Traffic sign:	8	Clouds:	3
		Traffic sign:	5	Road sing:	6		
		Clouds:	1	Clouds:	4		
Total	40	Road sing:	21	Traffic sign:	14	Clouds:	3
		Traffic sign:	14	Road sign:	12	Black object:	1
		Clouds:	4	Clouds:	8	Traffic sign:	1
		Highway:	1	Black object:	1		
				Asphalt:	1		
				Guardrail:	1		

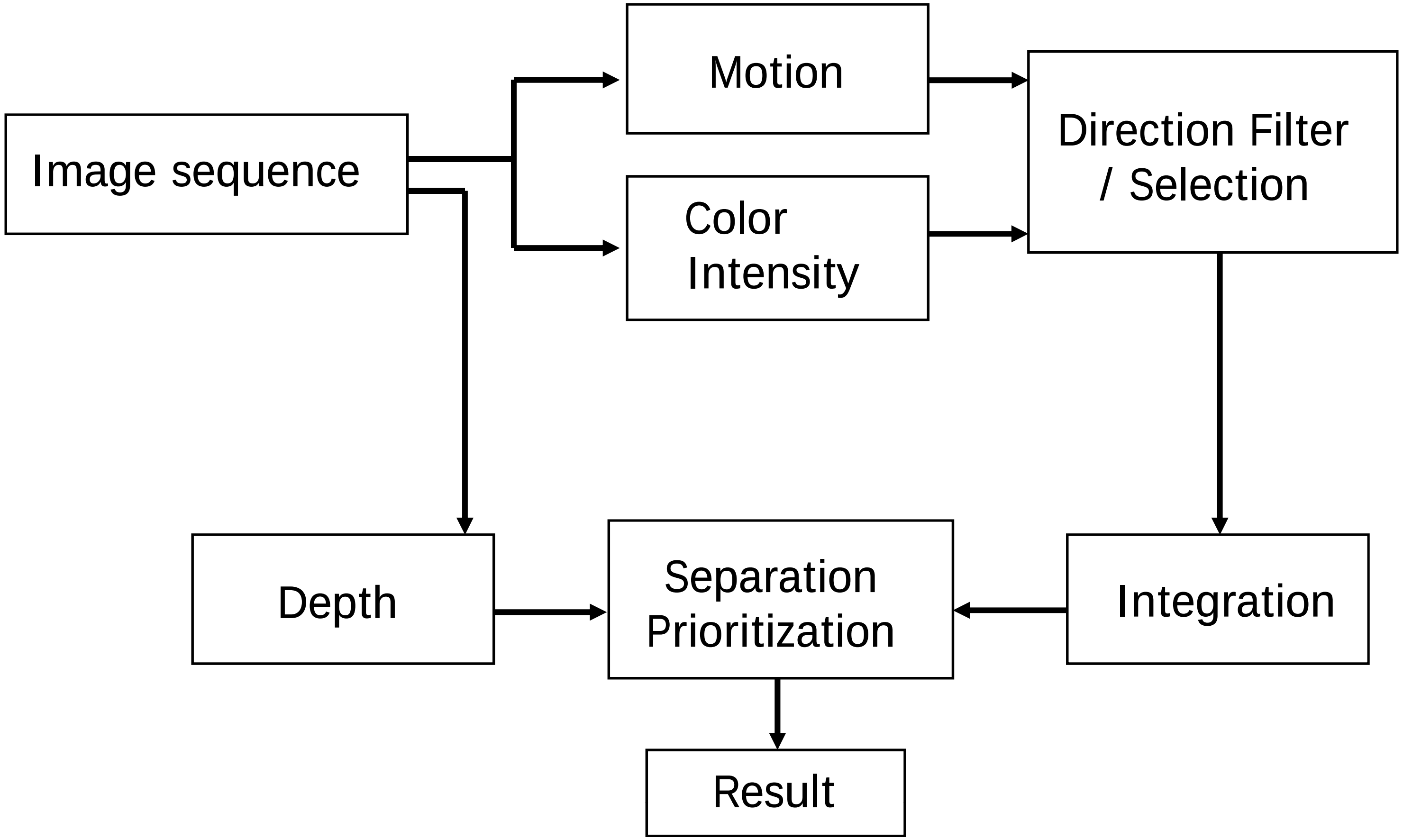
Proposed Method

- **Noise Elimination 2**
  - Use Flag Map
  - To prevent flickering (global-area noise) on time domain
    - at Frame(t), record/update state of motion existence to FlagMap
    - at t+1, refer to Flagmap(t)



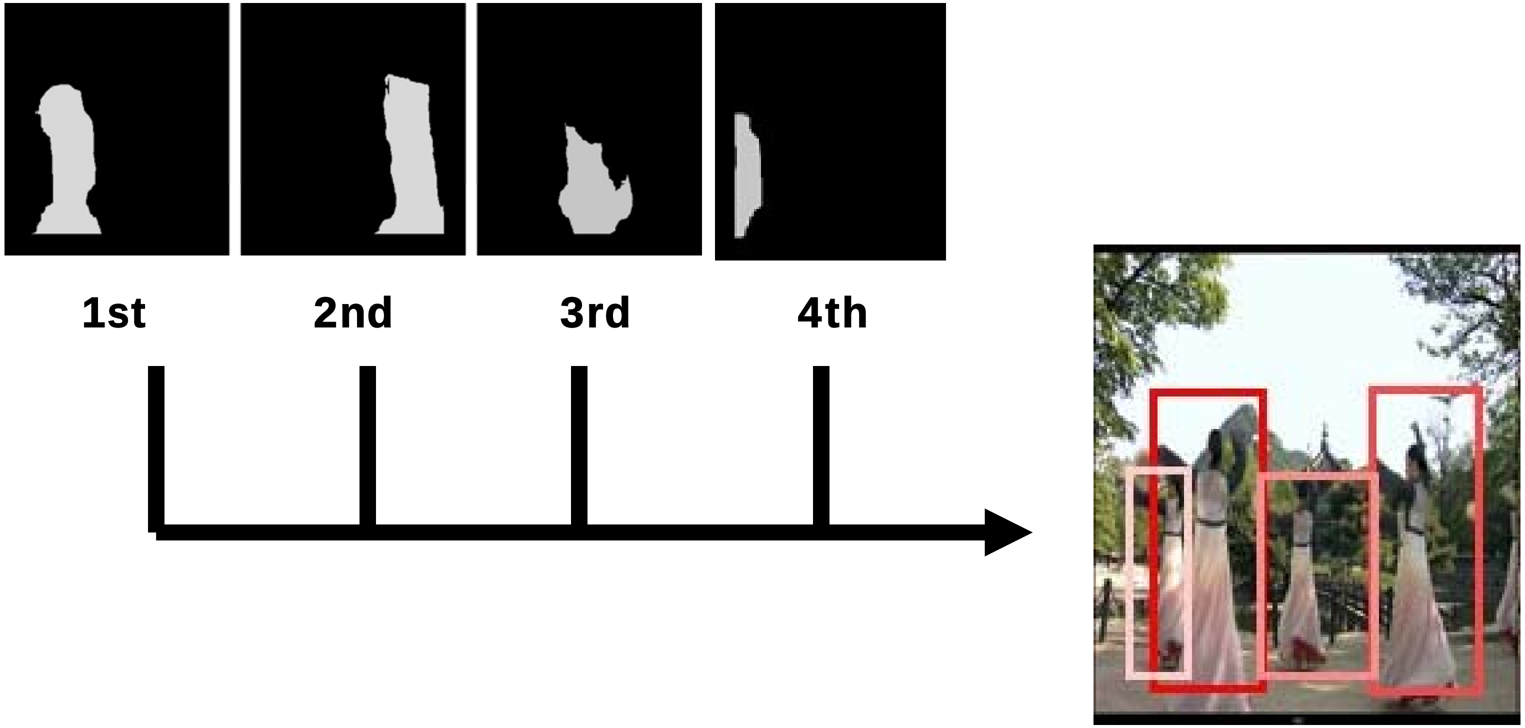
Proposed Method

- **Diagram**



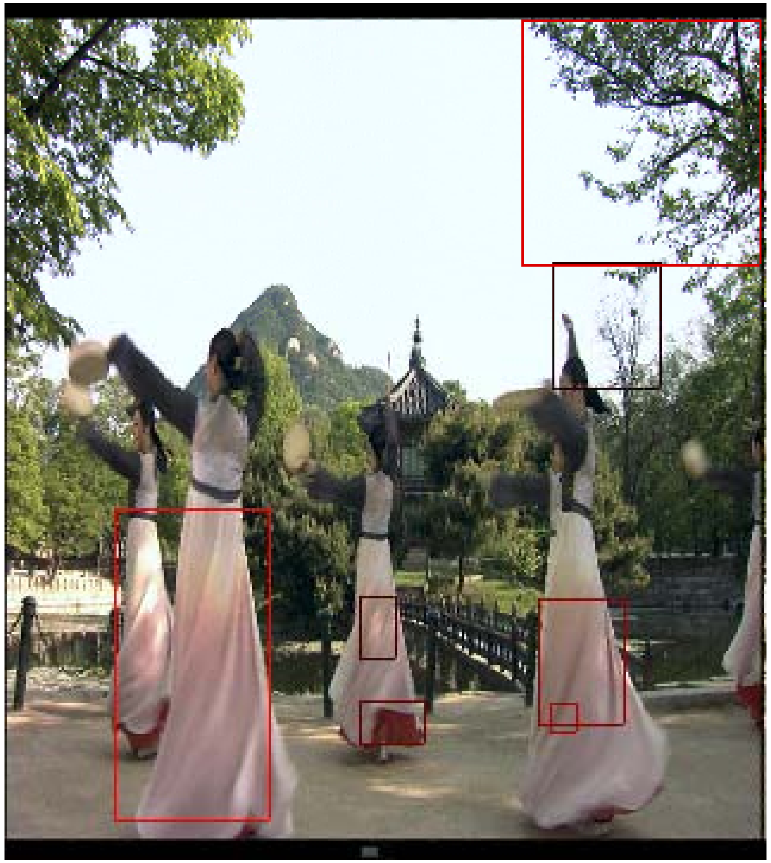
Experiments

- **Prioritization & Segmentation**
  - Separate and Prioritize each salient region
    - Based on depth value & saliency value

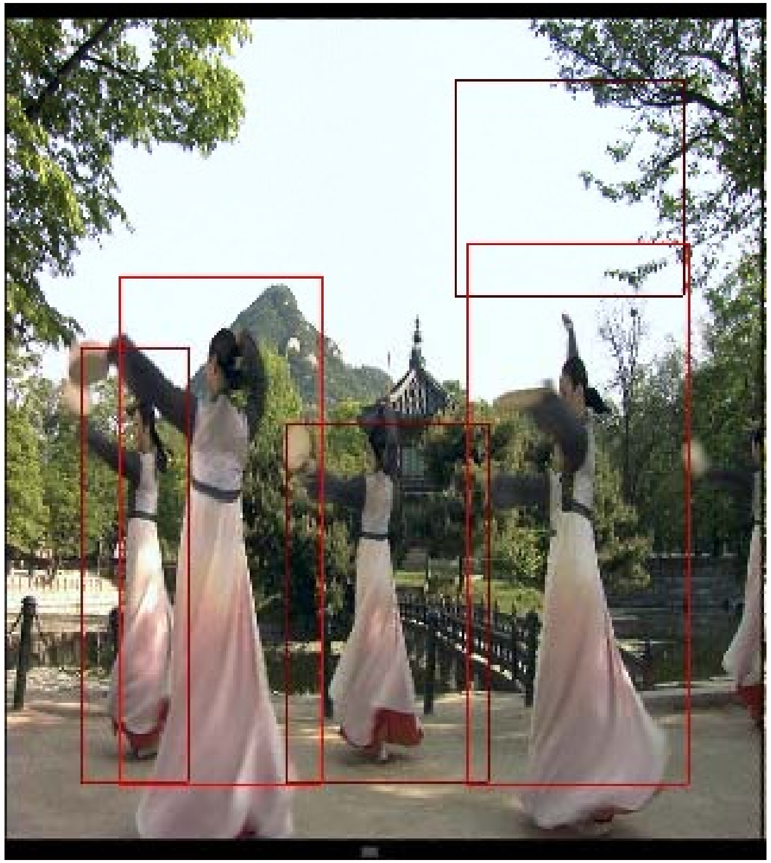


Experiments

- **Comparison**



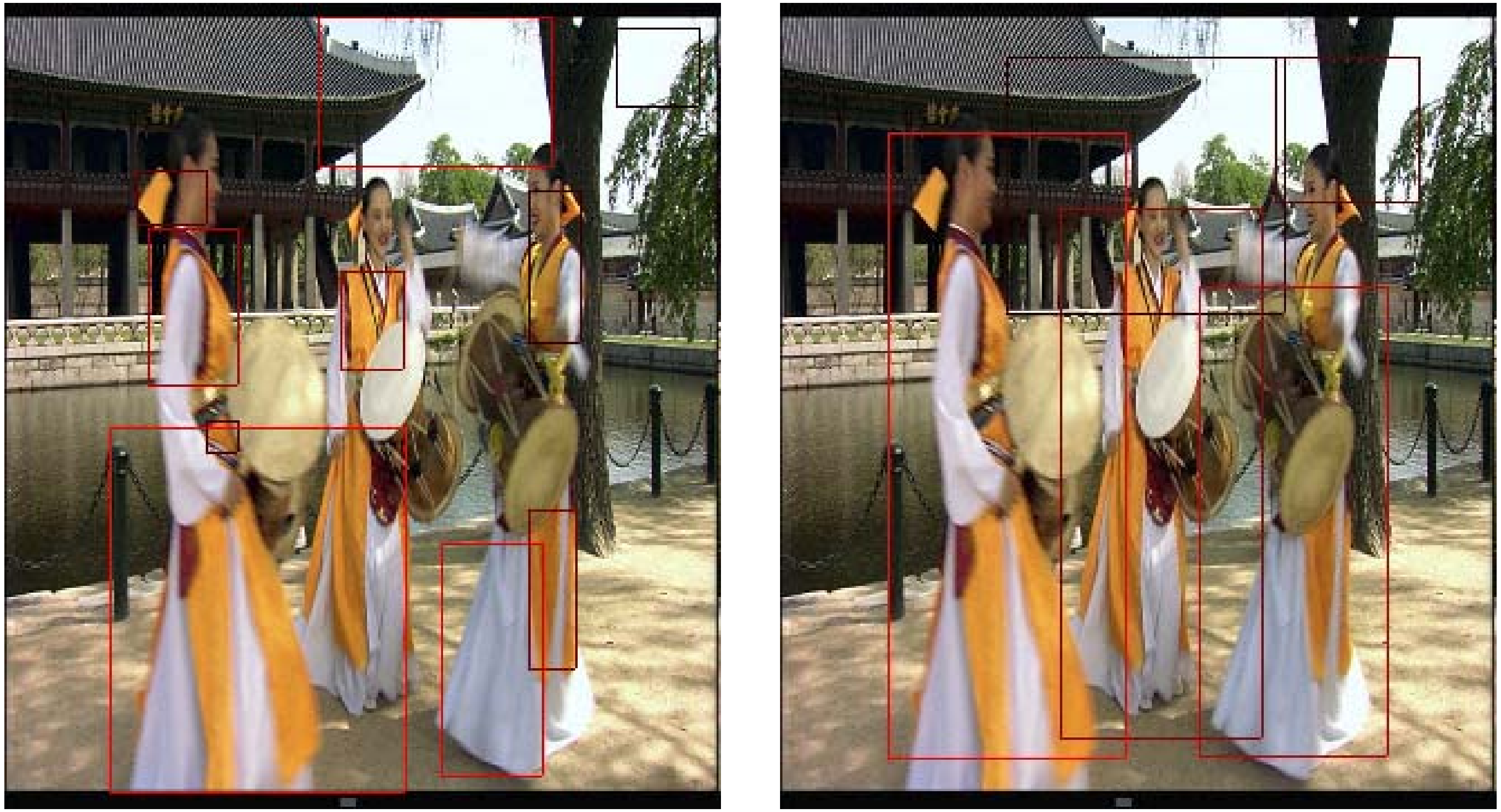
(a) Without Depth informtation



(b) With Depth information

Experiments

- **Comparison**



(a) Without Depth

(b) With Depth

Conclusion

- Generates a spatiotemporal salient map based on spatial & temporal features, and compensates for its inaccuracy using 3d depth information

- If some salient regions are partially occluded and have similar saliency value, each salient region is separated and prioritized sequentially

- To eliminate temporal noise and improve the accuracy of temporal saliency, psychological studies and Flag map are used. As a result of this, temporal saliency is obtained with higher accuracy

- Can not apply real time-processing system on high resolution image
  (computation for motion vector, stereo depth, convolution..)
  High accuracy / Low speed