AGENCE NATIONALE DE LA RECHERCHE ANR

i3S sophia antipolis

UNIVERSITÉ CÔTE D'AZUR

L3i Informatique Image Interaction Univ. La Rochelle

Content Based Image Retrieval based on implicit gaze annotations

Stéphanie Lopez

Director: Frédéric PRECIOSO

Co-advisors: Arnaud REVEL, Diane LINGRAND

Reviewers: Jenny BENOIS-PINEAU, Nicolas THOME

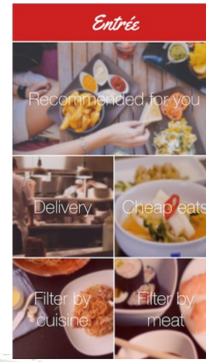
Examinators: Ebroul IZQUIERDO, Bernard MERIALDO

VISIIR

12/06/17

Content Based Image Retrieval

- Food image content

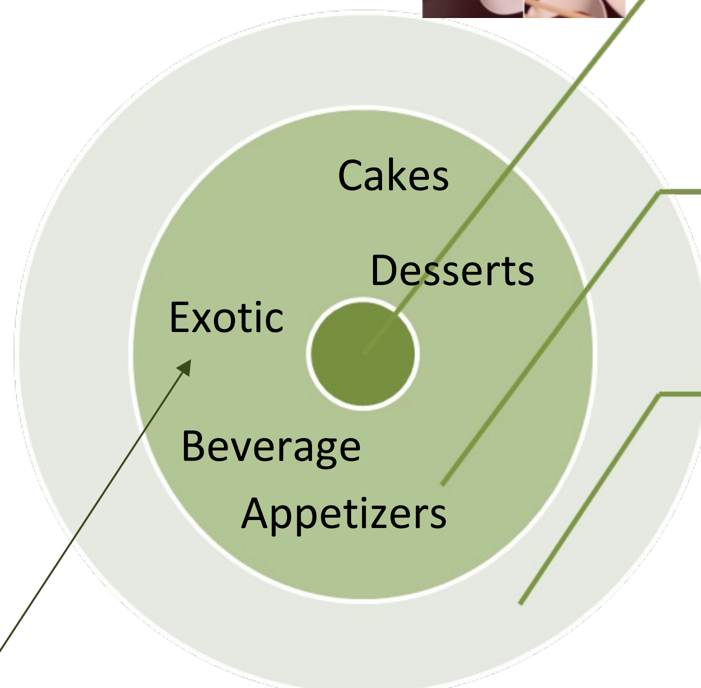


Need

What does **the user** want?

Food images

Images



Usually

What is it?
[Wang14]

Content Based Image Retrieval

Context

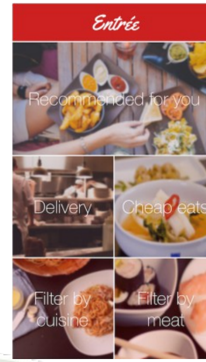
Existing works

1

2

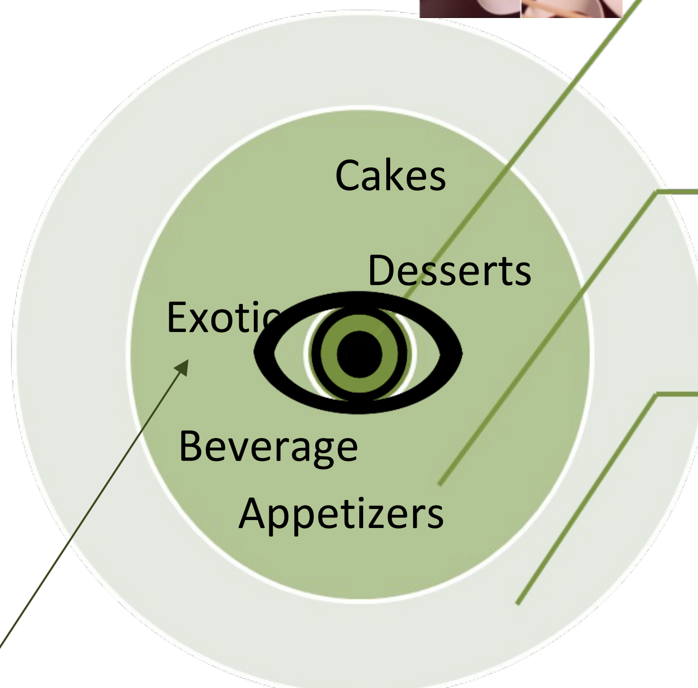
Conclusion

- Food image content



Need

What does **the user** want?



Food images

Images

Usually

What is it?
[Wang14]

Context

Existing works

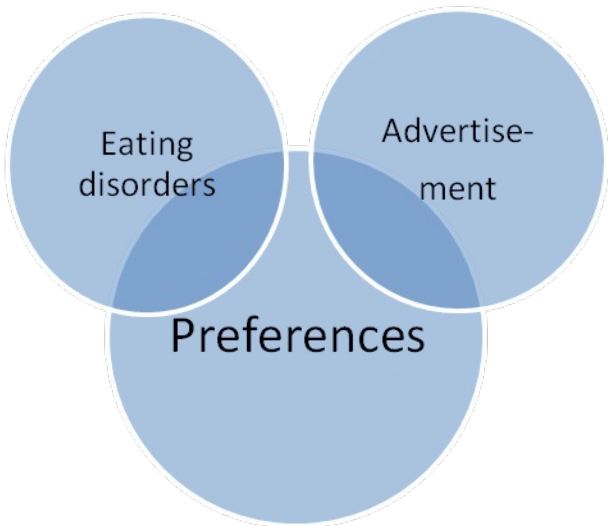
1

2

Conclusion

Visual Seek of Interactive Image Retrieval

Psychology (understand)



What does the user prefer? Why?

Computer Vision (model)

- Recognition
  - What is it?
- Retrieval
  - Which images correspond to?

How to alleviate the burden of annotation process?

Context

Existing works

1

2

Conclusion

Annotation

Human Perception

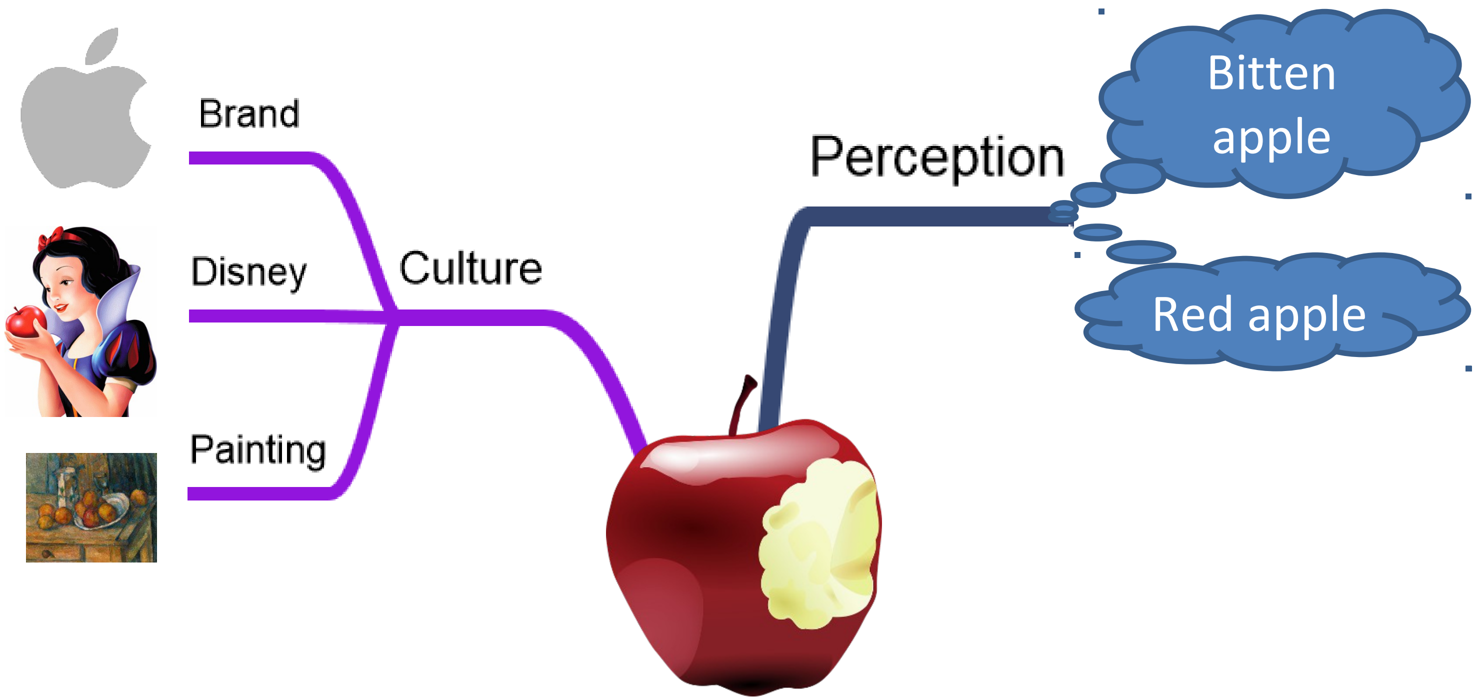
Context

Existing works

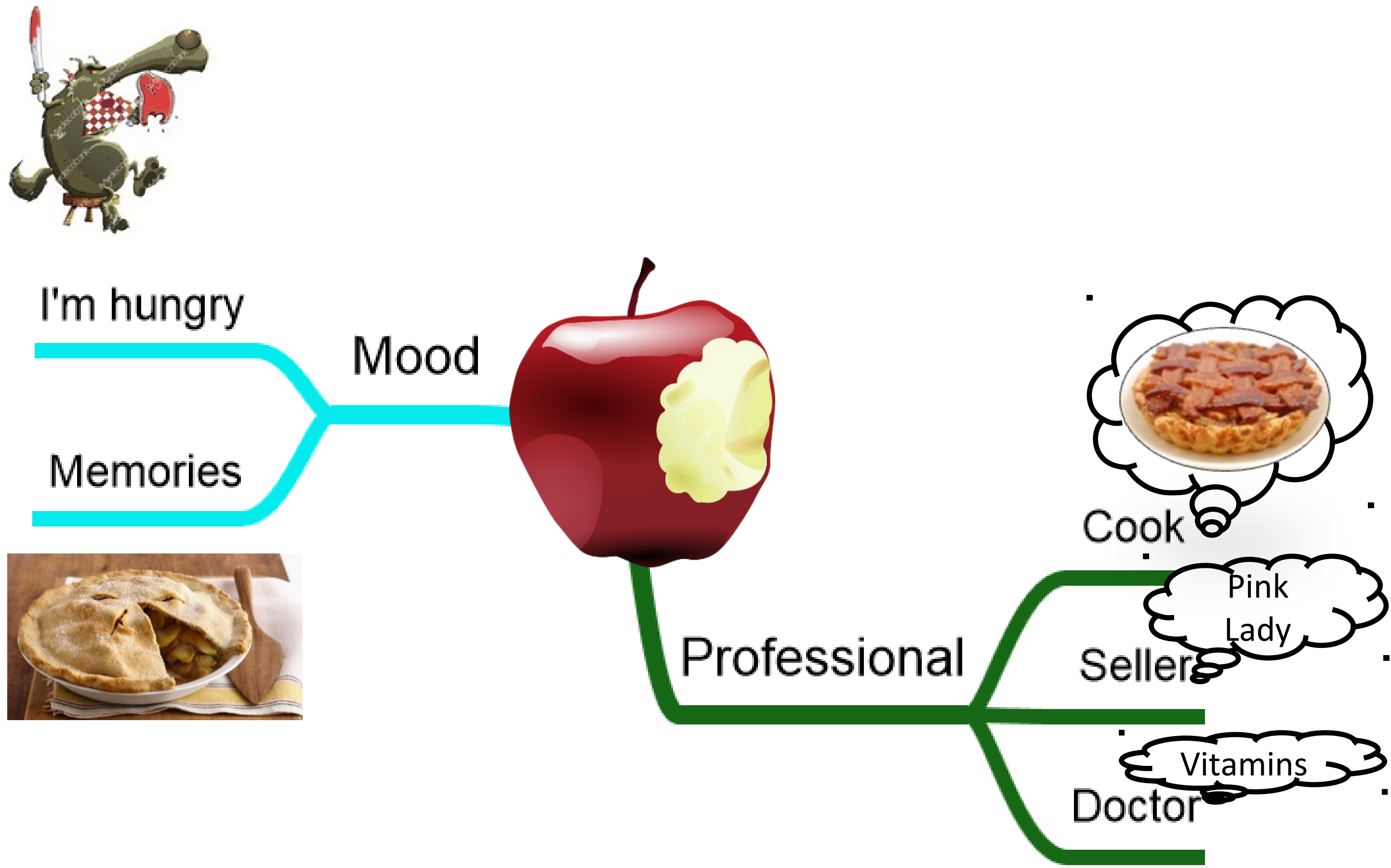
1

2

Conclusion



Memory / Details / Knowledge



Context

Existing works

1

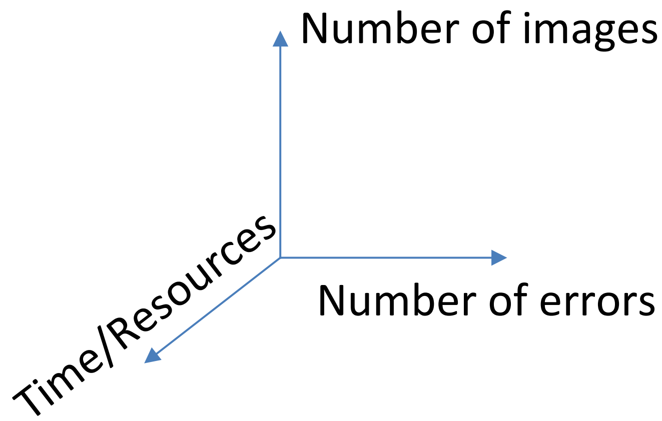
2

Conclusion

Constraints

Annotation constraints

- Annotations



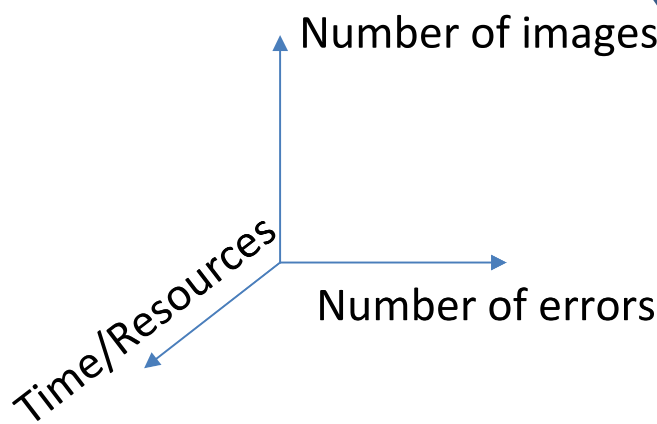
- Implicit annotations
  - Gaze distractors
  - Control [Jacob91]

Impact on classification

Constraints

Annotation constraints	Impact on classification

- Annotations



- Implicit annotations
  - Gaze distractors
  - Control [Jacob91]

- Few examples
- Weakly supervised learning
- Representation of the images relatively to a target category

Context

Existing works

1

2

Conclusion

Goals

Category identification by gaze

- Real time
- Gaze Based Intention Estimator (GBIE)
- Agnostic to users and categories

Towards user-centred concepts

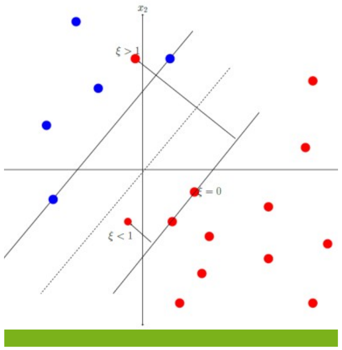
- Concept independence
- General / Subjective applications

Classification purpose

- Train with few examples (GBIE efficiency)
- Weakly supervised learning methods

  Measure on label reliability

Context | Existing works | Annotation | Classification | Conclusion



- Fundamentals about eye gazing
  - Gaze features
  - Gaze distractors
- Protocols in gaze studies
  - Psychological approaches
  - Computer vision

Eye tracking

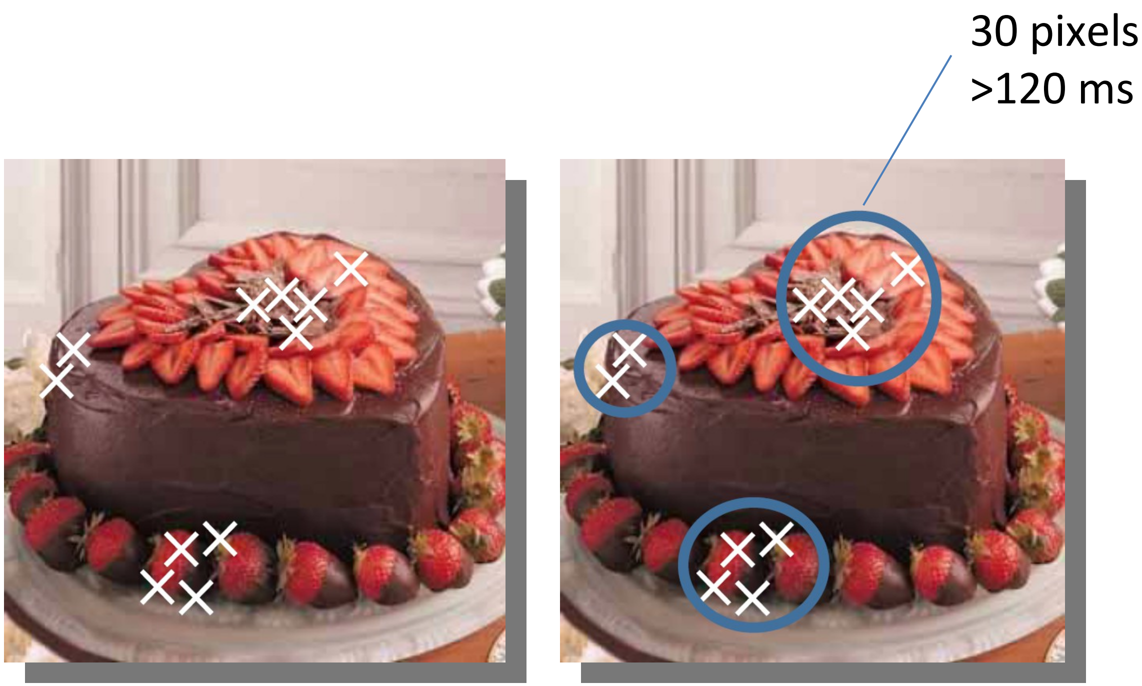
Eye tracker Tobii 32Hz – 60 Hz

- 2 infrared cameras
- Time sample
- Raw gaze data
  - X,Y position
  - Validity value
  - Pupil diameter
  - Timestamp

Gaze fundamentals

Micro-saccades -> fixations

Saccades



[Salvucci2000]
[Kozma2009]
[Auer2010]

Context

Existing works

1

2

Conclusion

30 Gaze Features

RAW FEATURES

- X,Y position
- First and last seen images
- Observation time
- Jumps between images

First seen image [Krajbich2010]

FIXATIONS

Number of fixations
•Per image
•First visit
•Last visit
Mean length
Position

Number of fixations [Kozma2009]

OTHERS

- Pupil diameter
- Gaze spread
- Gaze speed

Maximum pupil size [Hess1960]

Gaze distractors / Biases

Tasks
[Buswell35]

- Free viewing
- Emotion
- Memory
- Annotation

Feedbacks

- Mouse
- Keyboard
- Check boxes

Images
[Henderson13]

- Number
- Colors
- Texts
- Shape
- Size
- Background

Users
[de San Roman17]

- Boredom/ Tiredness
- Age
- Gender
- Culture
- Glasses
- Tastes
- Schedule
- Health [Castellanos09]

Context

Existing works

1

2

Conclusion

Protocols of gaze studies

Psychology

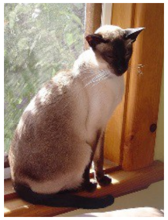
Preference

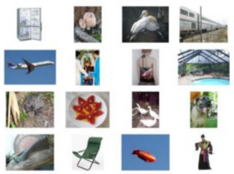
- Visual Preference Paradigm [Fantz58]

Computer Vision

Recognition

Retrieval





[Papadopoulos14]

[Hajimirza12]

Protocols of gaze studies

Psychology

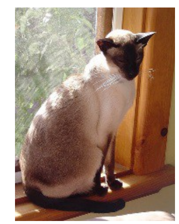
Preference

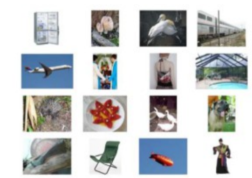
- Visual Preference Paradigm [Fantz58]

Computer Vision

Recognition

Retrieval





[Papadopoulos14]

[Hajimirza12]

Feedbacks

What is it?

Cat

What corresponds?

Is it a cat?

Is it interesting?

Content Based Image Retrieval (CBIR)

Context

Existing works

1

2

Conclusion

Gaze Annotation

- Understand human perception
  [Castellanos09]
- Automatize human annotation
  [Papadopoulos14]

Classification

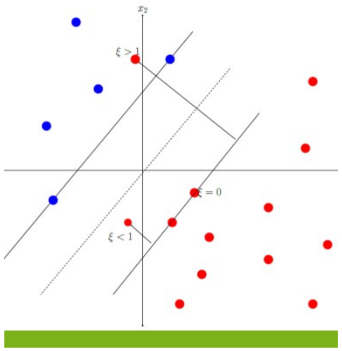
- Weakly supervised learning
- Representation of the images relatively to a target category

Combined strategy

Gazir

Gaze Based Intention Estimator

Context | Existing works | Annotation | Classification | Conclusion



Choice of our protocol

Visual Preference Paradigm

- **Intuitive**: I prefer this to that
  - Endless possibilities of identification contexts
  - Category independence
- **Limit specific gaze patterns**
  - User-independence
- **Real-time decision**
  - Not too many criteria of comparison
- **Display 2 images**
  - Smaller device
  - handfree decision

Context

Existing works

1

2

Conclusion

Goals reminder: Gaze Annotation

Concept identification by gaze

- Gaze Based Intention Estimator (GBIE)
- Real time
- Agnostic to users and categories

VPP

Towards user-centred concepts

- Concept independence
- General / Subjective applications

- Standard categories
- Food Specific categories
- Food General categories

Classification purpose

- Train with few examples
- Weakly supervised learning methods

Measure of uncertainty

- Task
  - Category identification
  - Image choices
  - Users' information
- Process

Category and User Independence

Limit the gaze habituation!

Standard categories	Food Specific categories	Food General categories
S1: Nice S2: La Rochelle	F1	F2
Animals Persons Vehicles Furniture	Beet salad Carpaccio Cannoli Ice cream	Appetizers Desserts Citrus Berries

Context

Existing works

1

- *protocol*
- GBIE

2

Conclusion

Image choice

Carpaccio

- Color
- Shape
- Size
- Limit gaze distractors
- No text
- Image difficulty [Tudor16]
  - ✓ Number of elements
  - ✓ Number of classes
  - ✓ Confusability of the elements

Users' information

Questionnaire

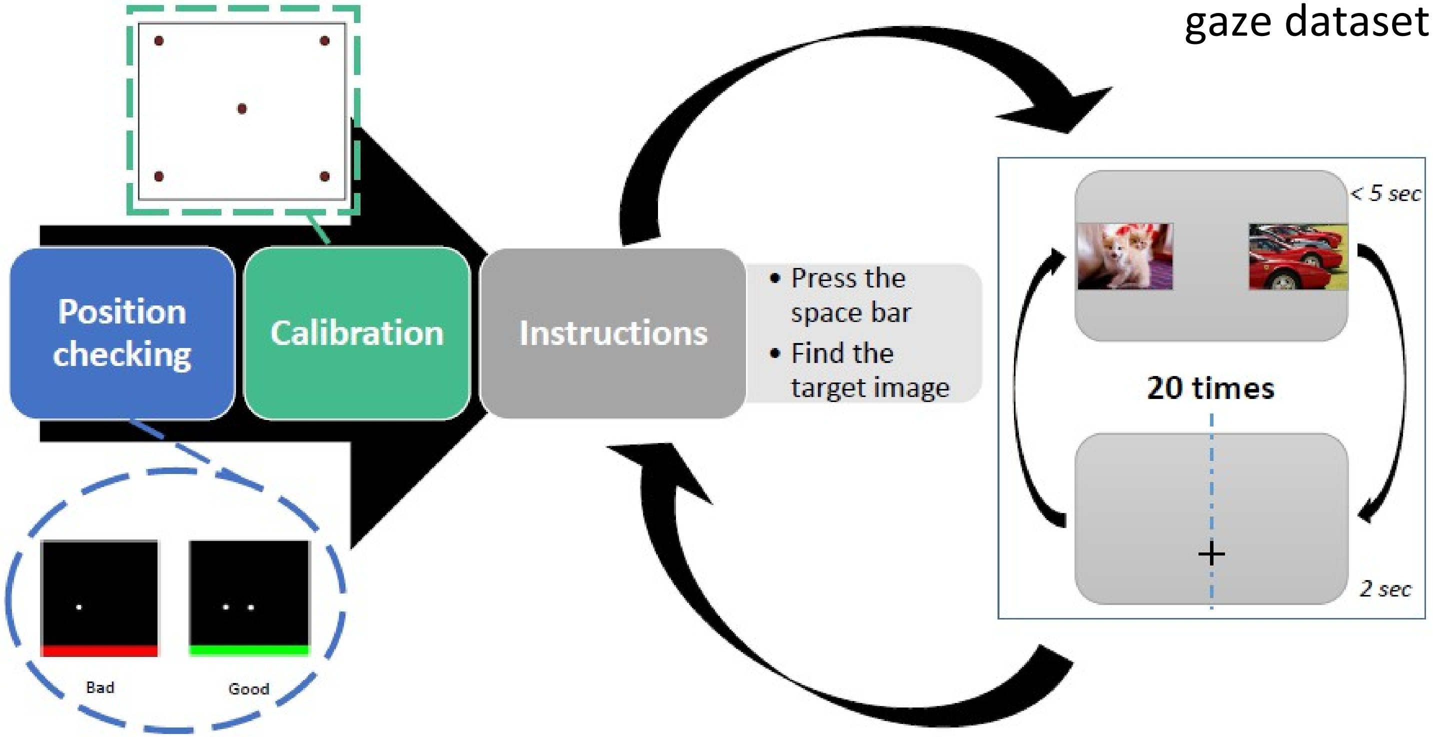
- Age
- Gender
- Schedule
- Glasses
- Tastes

	S1	S2	F1	F2
	40	46	46	52
	1.7 sec	1.9 sec	2.0 sec	2.2 sec

[Duchowski02] 400 to 800ms to understand the content => predict in <1sec ?

Process

Publicly available gaze dataset



- Most discriminant feature among 30 features
  – User independent
  – Category independent
- Real-time constraint
- Gaze Based Intention Estimator (GBIE)
- Comparison to existing works

Context

Existing works

1

- protocol
- *GBIE*

2

Conclusion

Most discriminant gaze feature

- **Hypothesis**:
  - The simpler, the more generalizable
    - User
    - Target category
- **Question**:
  - Is one feature enough to infer the user's choice?
- **Challenges**:
  - No correlation with visualisation time
  - Method that provides a measure of uncertainty
- **User independence** (for a given experiment)
  – Merged gaze data of participants

Most discriminant feature

Context

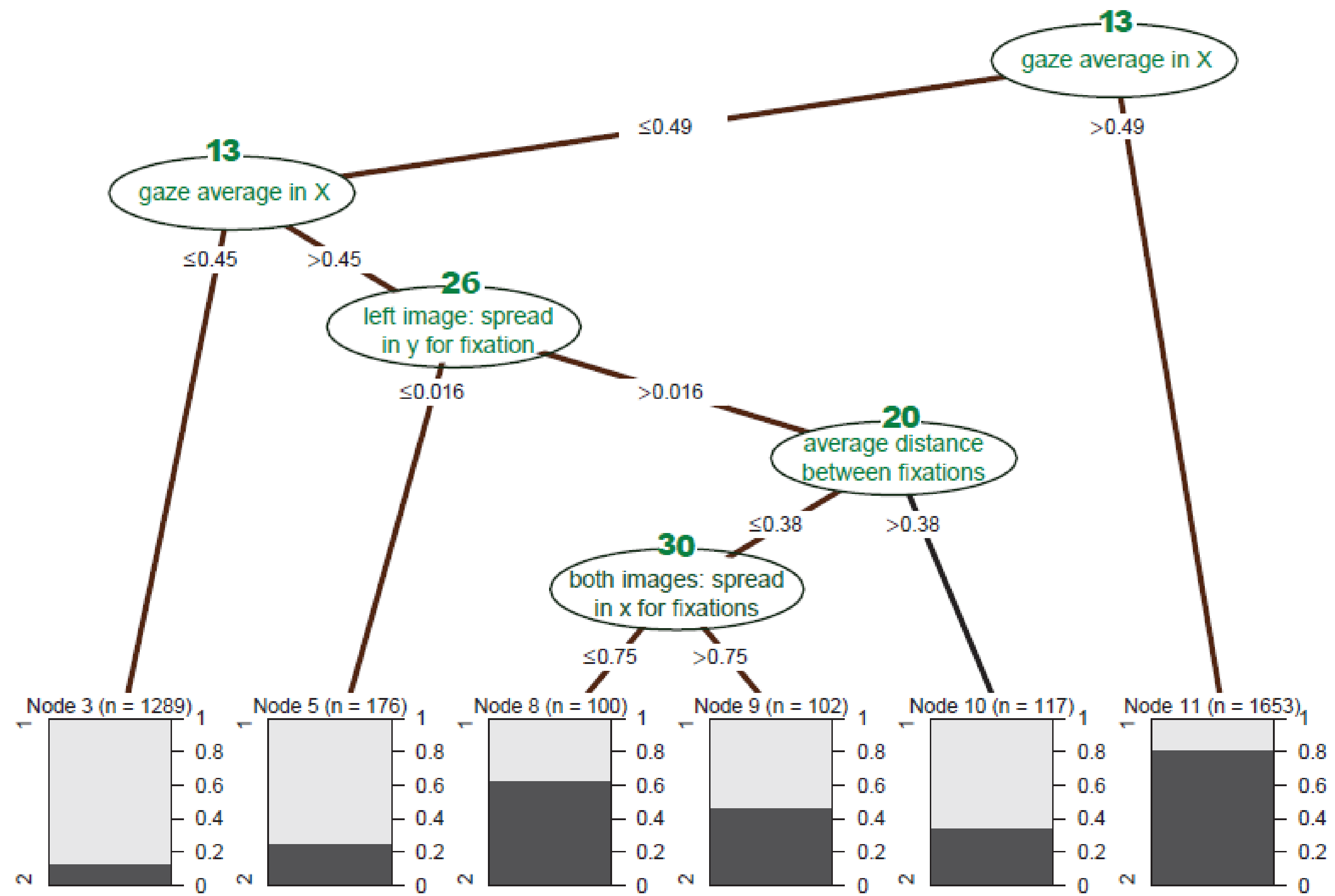
Existing works

1

- protocol
- *GBIE*

2

Conclusion



Target image	left	left	right	left	left	right
Confidence	90%	78%	61%	55%	67%	79%

Context

Existing works

1

- protocol
- *GBIE*

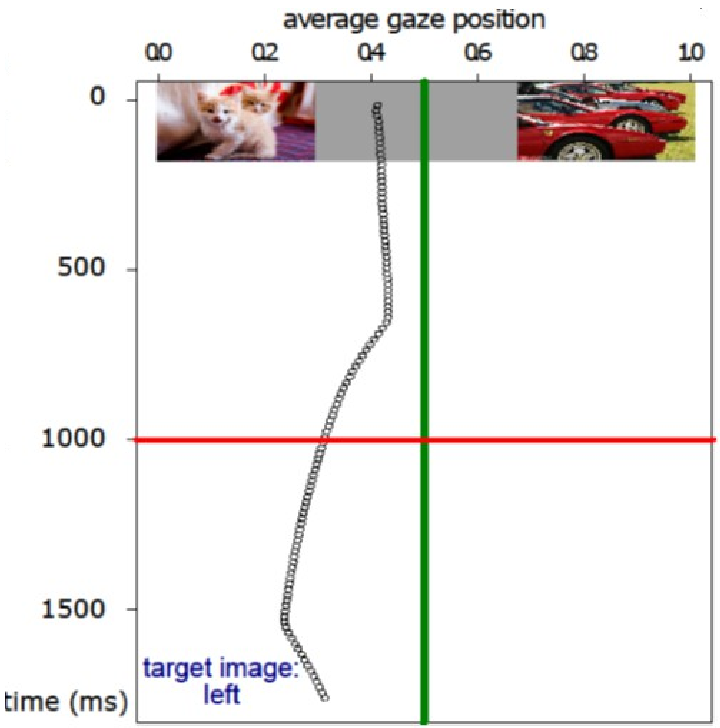
2

Conclusion

Most discriminant feature

Average horizontal gaze position

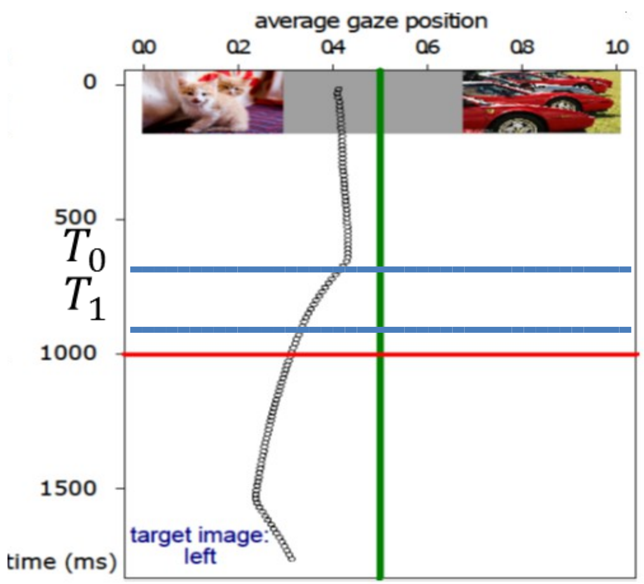
ANIMALS



- Category independence
  - Same feature for all the experiments

Real-time constraint

Not until the end of the visualisation time

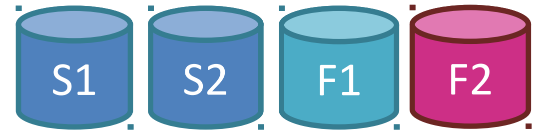


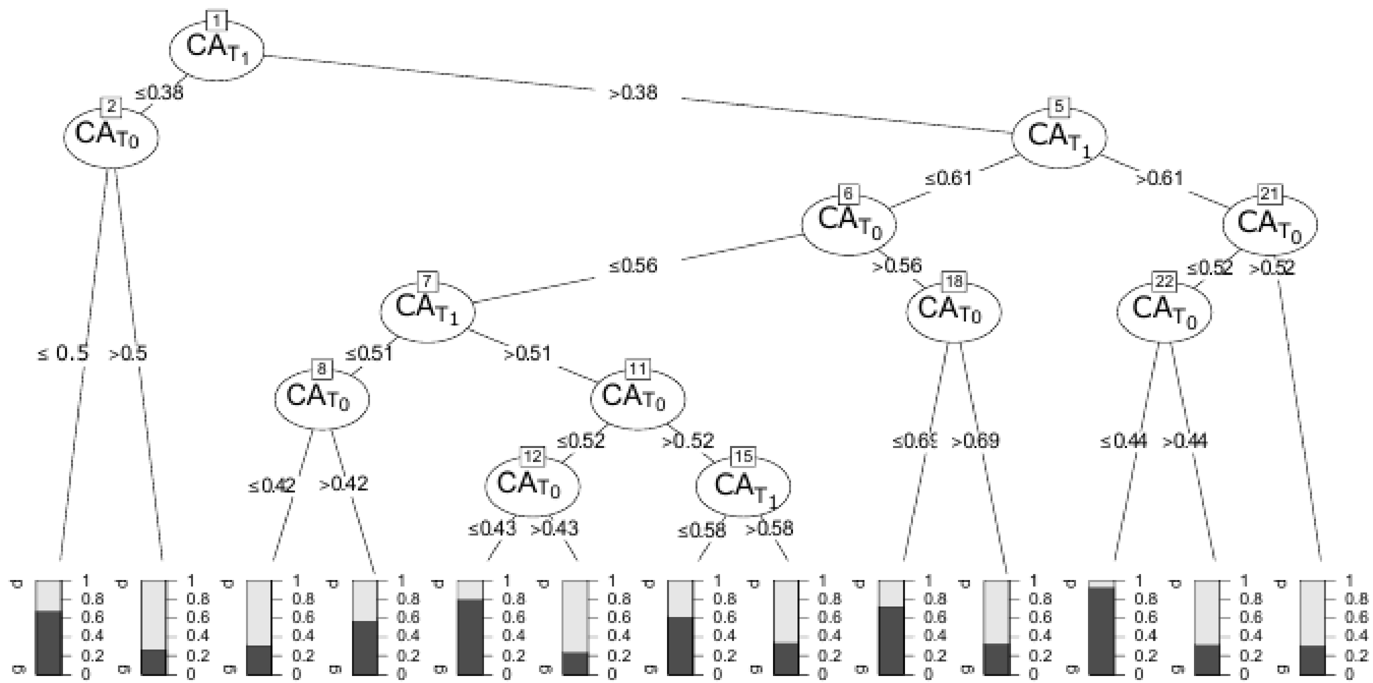
- Cumulative average of horizontal gaze position at
  - T_0
  - T_1

T0 (ms)	640		672		768	800	832
T1 (ms)	800	960	832	992	928	960	992

Real-time constraint

- One decision tree per experiment





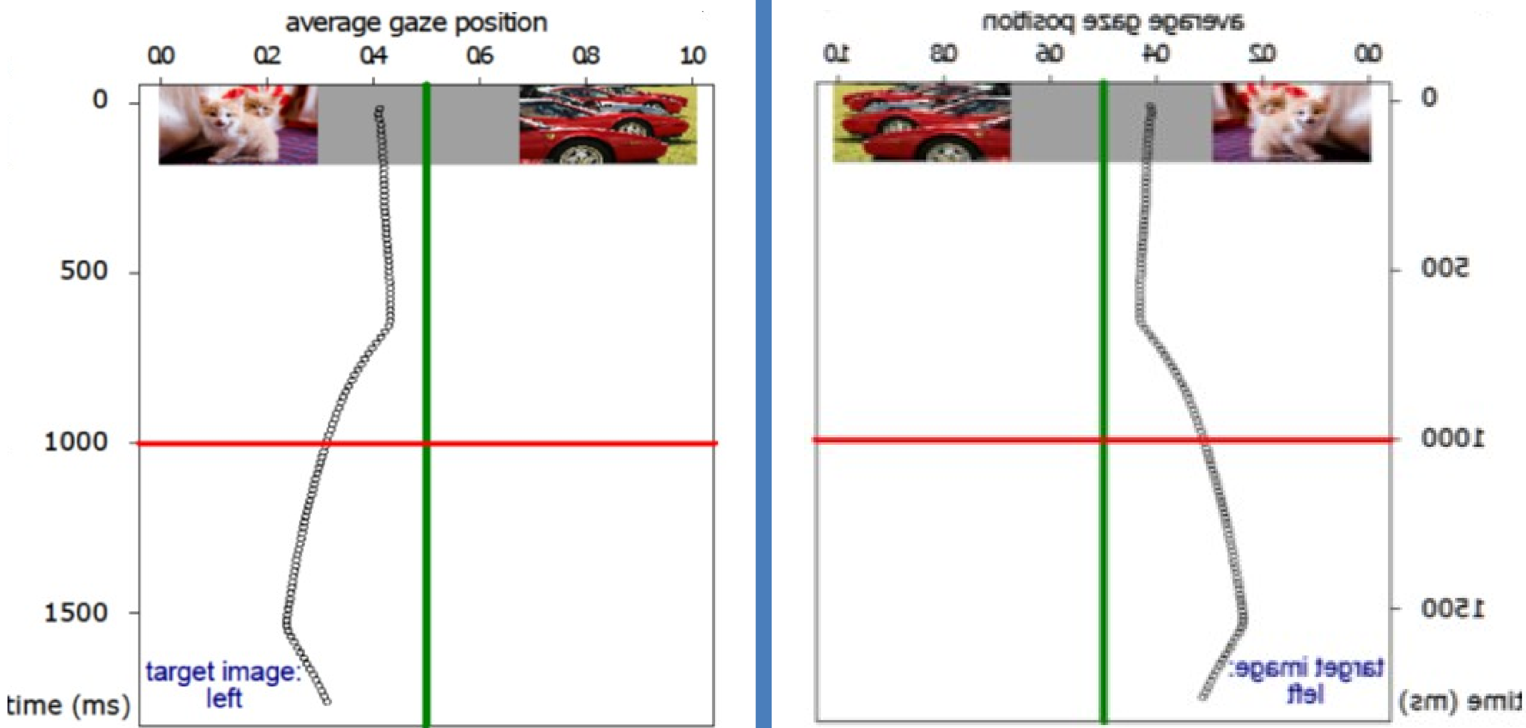
4 possible GBIE

Validity of the prediction

	T0=800ms, T1=960ms
S1	67.8%
S2	63.6%
F1	81.1%
F2	54.4%

Mirror data

Symmetrical property



Selection and validation of the GBIE

S1, S2, F1: experiments

Sym S1, Sym S2, Sym F1: mirror data

<table>
<tr><th></th><th colspan="2">S1</th><th colspan="2">F1</th></tr>
<tr><td>S1</td><td>(67.8%)</td><td rowspan="4">User independent</td><td>55.2%</td><td rowspan="4">Not category independent</td></tr>
<tr><td>Sym S1</td><td>64.0%</td><td>54.8%</td></tr>
<tr><td>S2</td><td>63.3%</td><td>53.3%</td></tr>
<tr><td>Sym S2</td><td>60.5%</td><td>53.3%</td></tr>
<tr><td>F1</td><td>71.5%</td><td rowspan="2">Category independent</td><td>(81.0%)</td><td rowspan="2">User independent??</td></tr>
<tr><td>Sym F1</td><td>64.4%</td><td>81.0%</td></tr>
</table>

Context

Existing works

1

- protocol
- *GBIE*

2

Conclusion

Selection and validation of the GBIE

S1, S2, F1: experiments

Sym S1, Sym S2, Sym F1: mirror data

	S1		F1	
S1	**(67.8%)**	User independent	55.2%	Not category independent
Sym S1	64.0%		54.8%	
S2	63.3%		53.3%	
Sym S2	60.5%		53.3%	
F1	71.5%	Category independent	**(81.0%)**	User independent??
Sym F1	64.4%		81.0%	

GBIE built on gaze data of S1
Average validity: ~70%

Context

Existing works

1

- protocol
- *GBIE*

2

Conclusion

Comparison with existing works

Until the end of the visualisation time

	GBIE	[Kozma2009] Max. number of fixations	[Hess1960] Max. size of pupil	[Krajbich2010] First seen image
S1	-	51.0%	**67.7%**	49.4%
S2	-	22.5%	**65.3%**	50.2%
F1	-	57.1%	60.2%	**85.1%**

Context

Existing works

1

- protocol
- *GBIE*

2

Conclusion

Comparison with existing works

Until 960 ms

	GBIE	[Kozma2009] Max. number of fixations	[Hess1960] Max. size of pupil	[Krajbich2010] First seen image
S1	**67.8%**	37.2%	55.8%	48.9%
S2	**63.3%**	34.8%	58.4%	50.1%
F1	71.5%	59.8%	65.0%	**83.6%**

Context

Existing works

1

2

Conclusion

Goals reminder

Concept identification by gaze

- ✓ Gaze Based Intention Estimator (GBIE)
- ✓ Real time
- ✓ Agnostic to users and categories

- ICIP2015
- publicly available gaze dataset

Towards user-centred approaches

- ✓ Concept independence
- ? General / Subjective applications

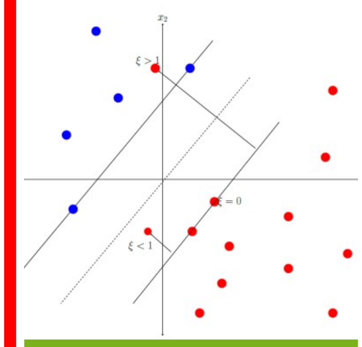
AVI2016

Classification purpose

- ➢ Train with few examples
- ➢ Weakly supervised learning methods

Classify with a traning set annotated with our GBIE

Context | Existing works | Annotation | Classification | Conclusion



Context

Existing works

1

2

- *Methods*
- Results

Conclusion

- Dataset (standard categories)
- Standard Classification
- Handling label uncertainty
- Strategies
  - Criteria of label discrimination
  - 2 contexts of classification

Data

- **VOC2007** (20 subcategories -> 4 general categories)
  - Training: 5011 images

Bird		Aeroplane	Bottle
Cat		Bicycle	Chair
Cow		Boat	Dining table
Horse		Bus	Potted plant
Dog		Car	Sofa
Sheep		Motorbike	TV
		Train	

Animals | Persons | Vehicles | Furniture

General categories

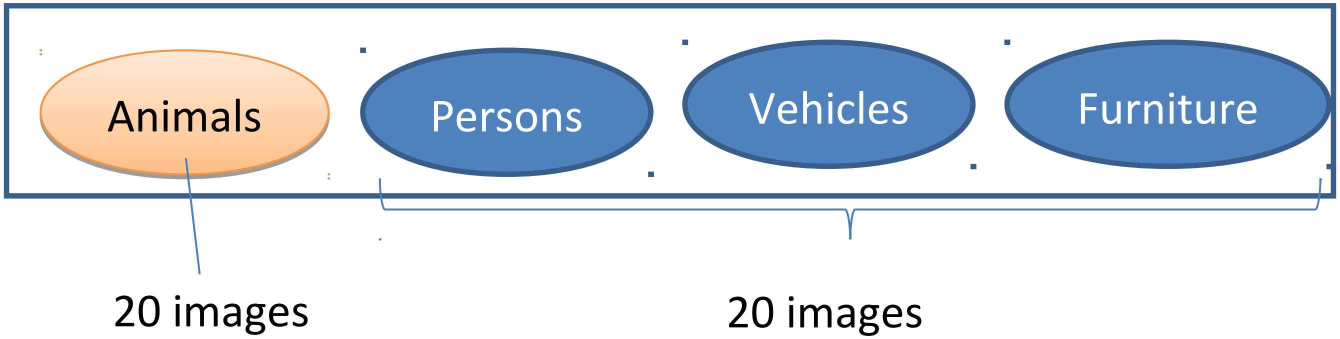
- Deep features provided by [Durand15]
- Test set: 4952 images

→Is it possible to classify images according to general categories?

- TC5011

Data

- **<u>VOC2007</u>** (20 subcategories -> 4 general categories)
  - Training: 5011 images
  - Test: 4952 images
- **<u>Our training set</u>**:
  - 40 images seen by 86 participants (S1 + S2)
    - 20 targeted images
    - 20 untargeted images



Small training sets and Uncertain labels

- Deep features provided by [Durand15]
- Test set: 4952 images

→Is it possible to classify images according to general categories?

- TC5011

C-SVM

→**If so,** is it possible to classify images with a training set that is 100-time smaller and randomly selected examples?

- TC40

C-SVM

→**If so,** is it possible to classify images with uncertain labels?

- GBIE40

C-SVM / handling label noise

Performance metrics

Accuracy

- Classification

(Computer Vision)

- The whole dataset

Precision @k

- Retrieval

(User-centred)

- The first **k** most relevant images

Is C-SVM robust to label noise or should we use other methods?

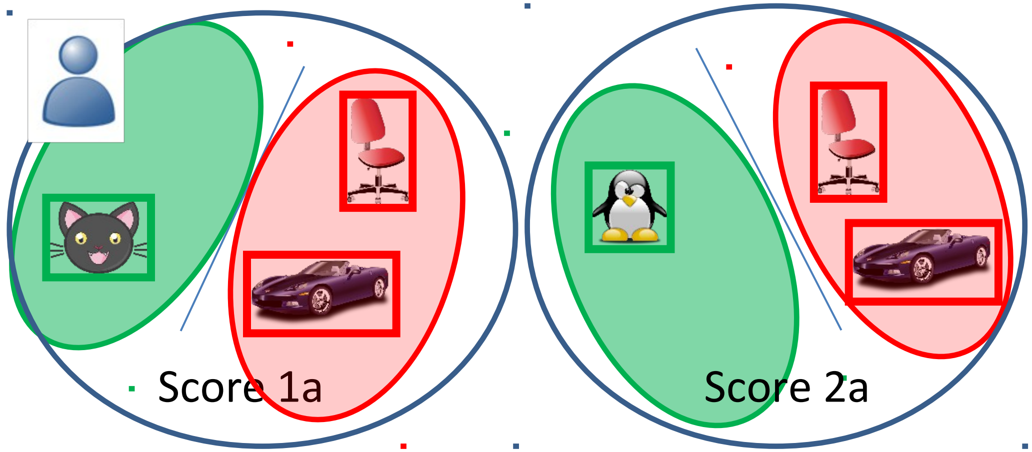
Label Uncertainty Survey [Frenay14]

- **Robust** [Teng01]
  - Is C-SVM robust to label noise
  - powerSVM [Zhang2012]

Representativeness (powerSVM)

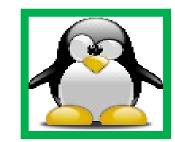
Representativeness score = classifier score of 1 + example vs all - examples

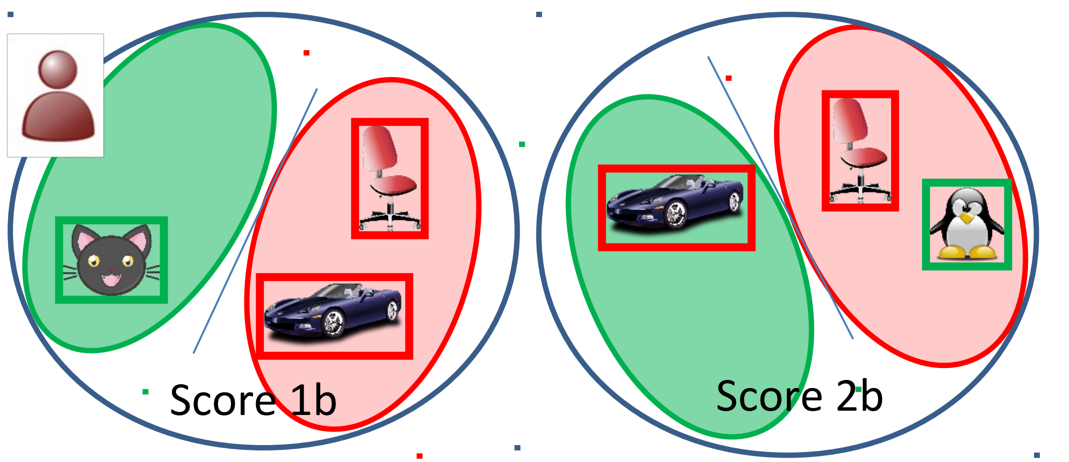
<u>for EACH positive example</u> => RANK



Expectations

Less/more reliable than





Expectations

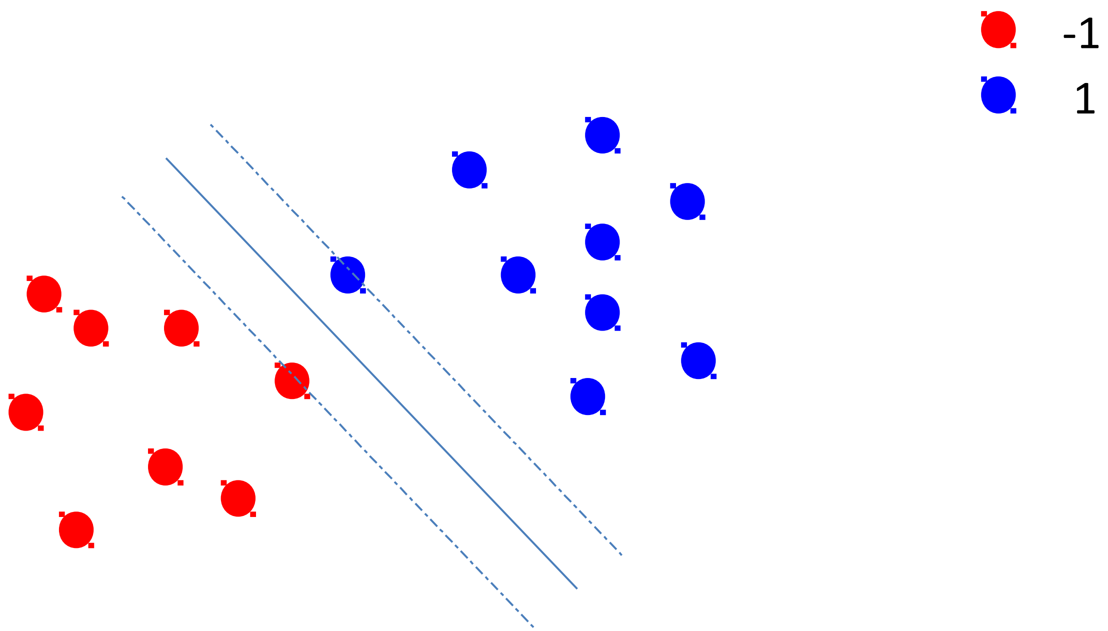
Less reliable than

Label Uncertainty Survey [Frenay14]

- **Robust** [Teng01]
  - Is C-SVM robust to label noise?
  - powerSVM
- **Data cleansing** [Garcia15]
  - Not possible with small training sets
- **Label noise tolerant** [Niaf14]
  - P(robabilistic)-SVM

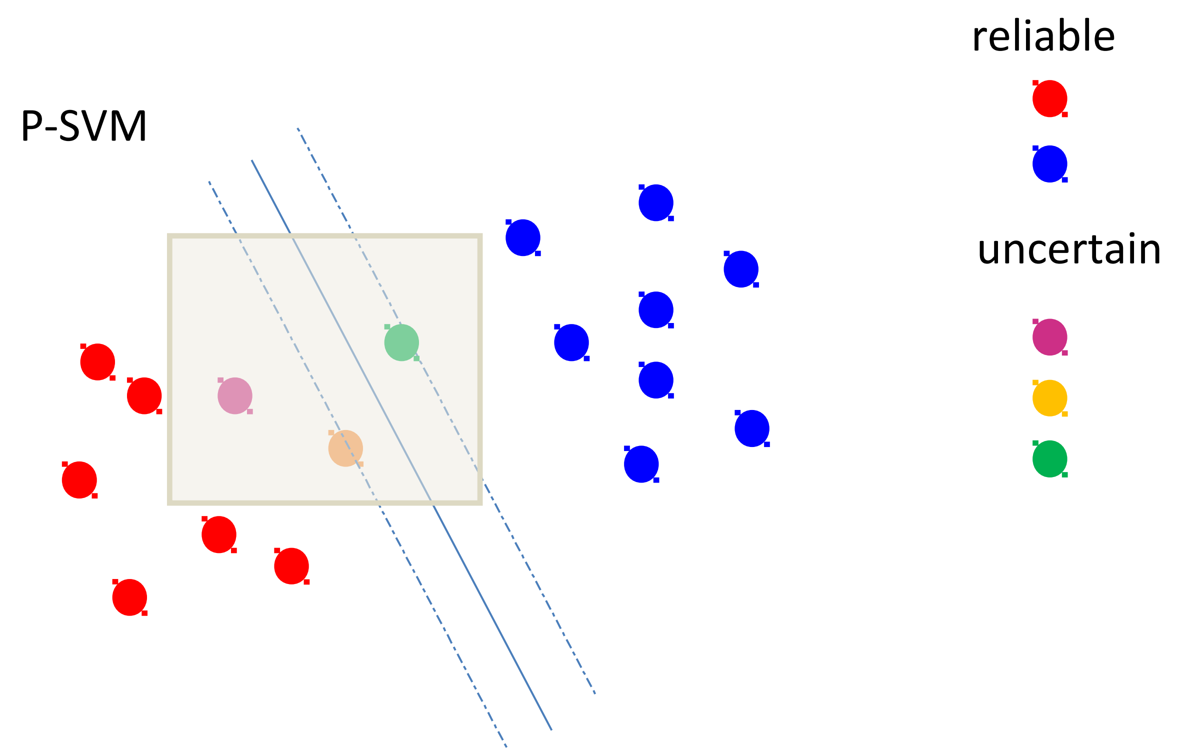
P-SVM derived from C-SVM

Find the hyperplane that maximizes the margin



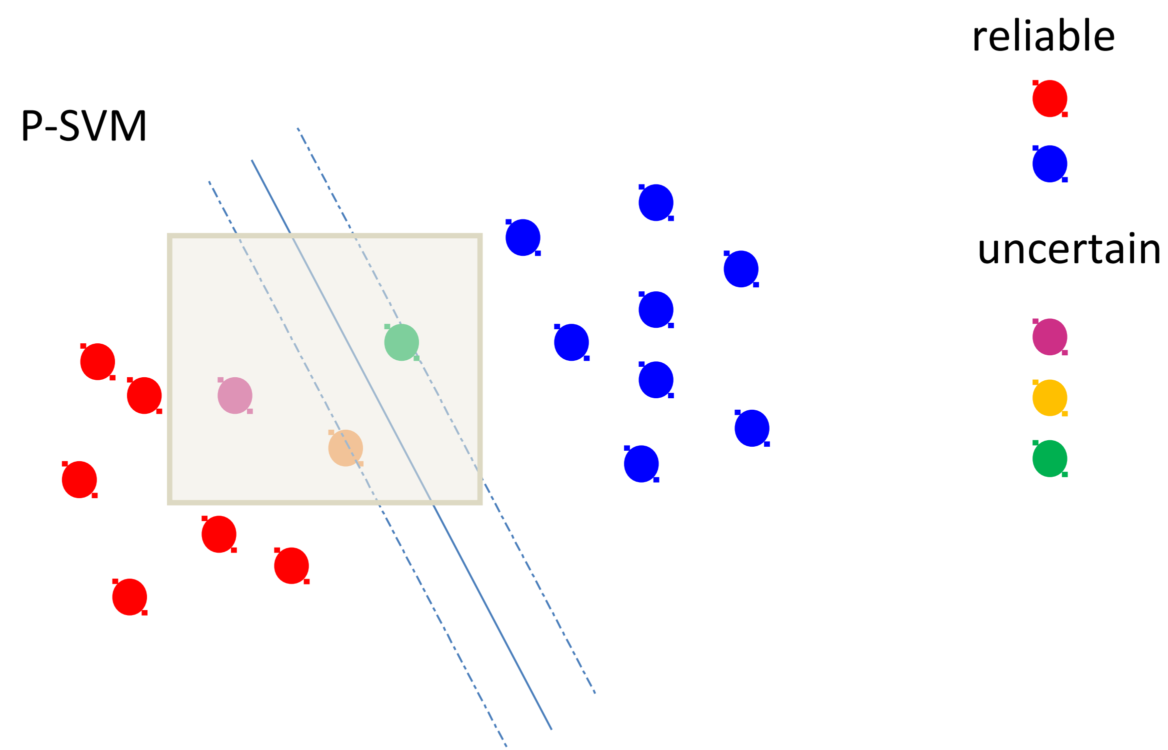
Handling uncertain labels with P-SVM

Hyperplane calculated with C-SVM and refine with C-SVR



Handling uncertain labels with P-SVM

Hyperplane calculated with C-SVM and refine with C-SVR



All the gaze labels are uncertain.
How can we identify images with the MOST RELIABLE labels?

Context

Existing works

1

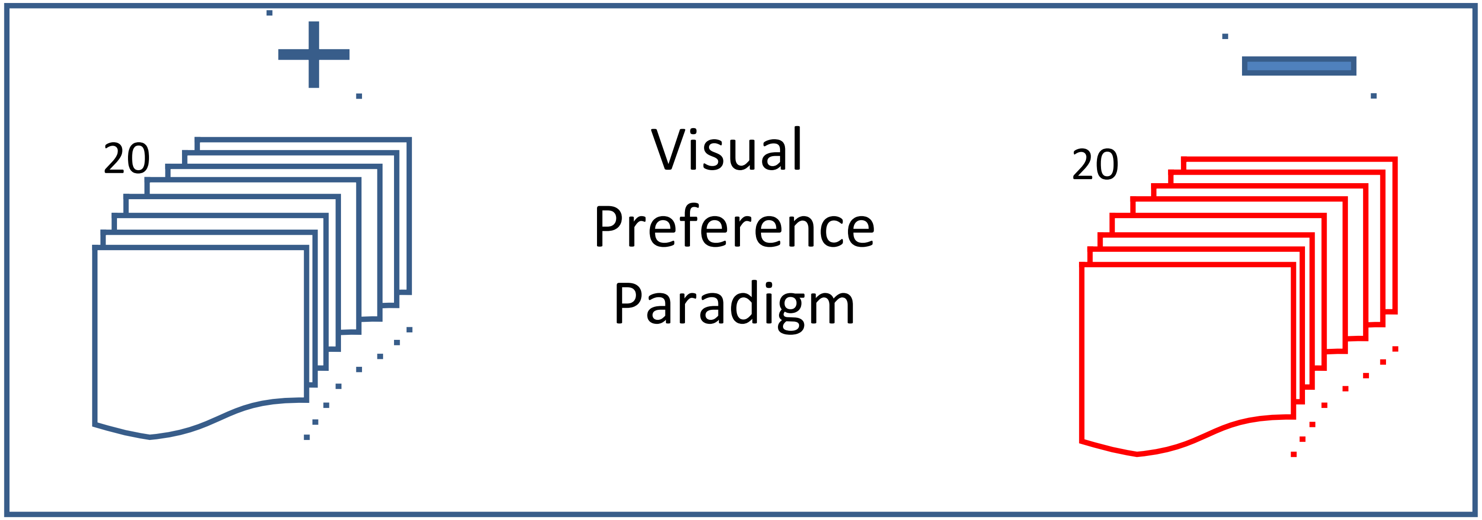
2

- *Methods*
- Results

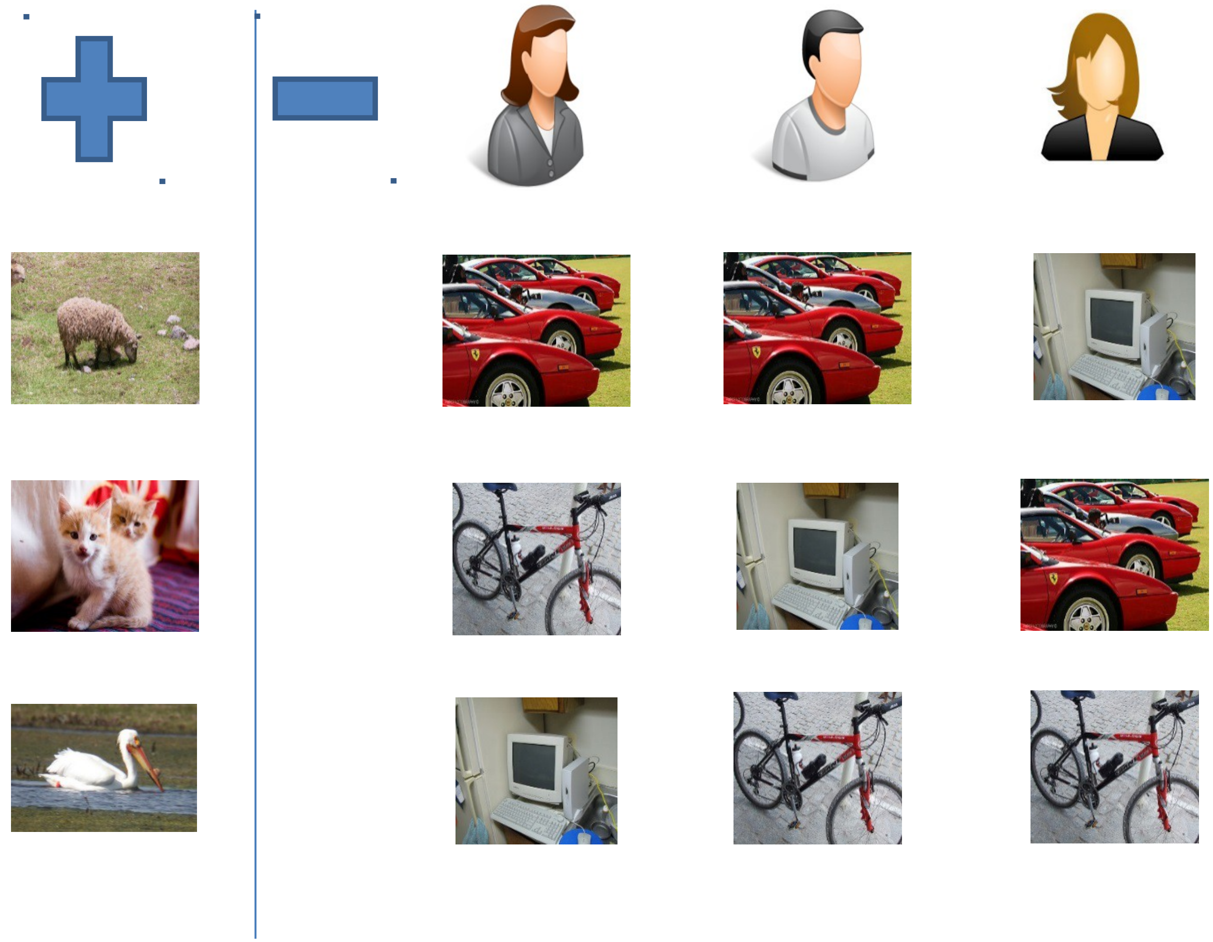
Conclusion

Reliability discrimination

- Discriminate the most reliable from the most uncertain (positive class)
  - Most reliable -> class labels (-1,+1)
  - Most uncertain -> probabilistic labels (p_i=P(y_i=1|x_i))

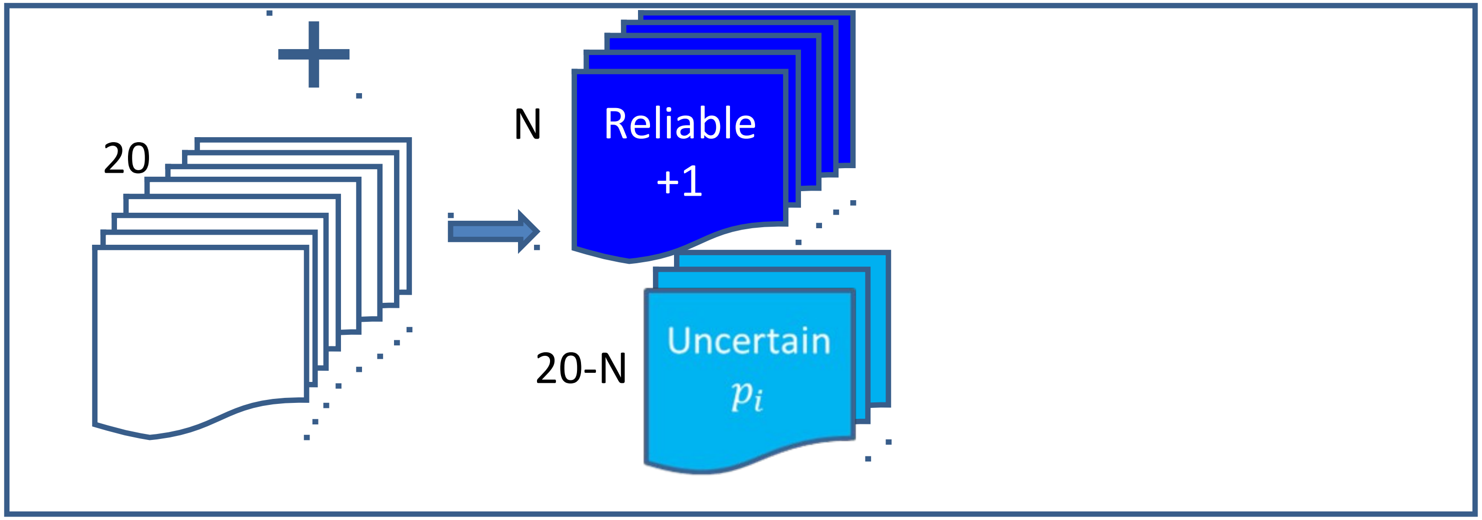


Reminder VPP



Reliability discrimination

- Discrimination on positive images
  - Majority vote
  - Representativeness of the images relatively to the category [Zhang12]



Label discrimination

Majority score

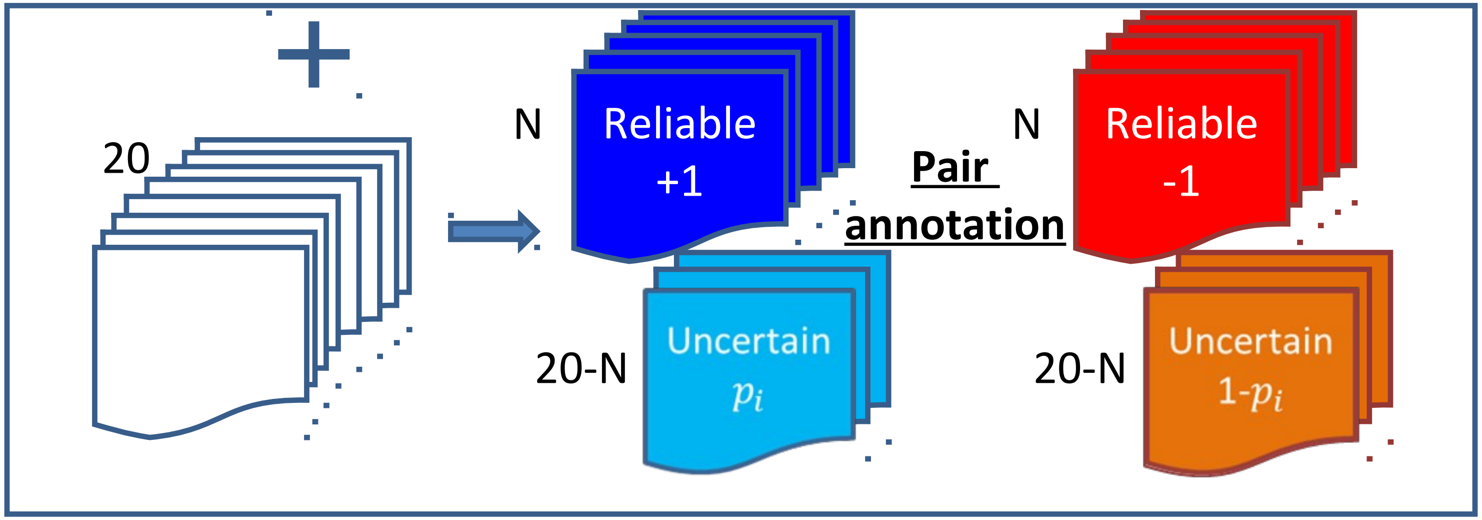
- Calculate how many participants have assigned a positive label
- Select N positive images with the highest number of positive labels

Representativeness score

- Calculate 20 scores of representativeness
- Select the N images with the highest scores

Label discrimination

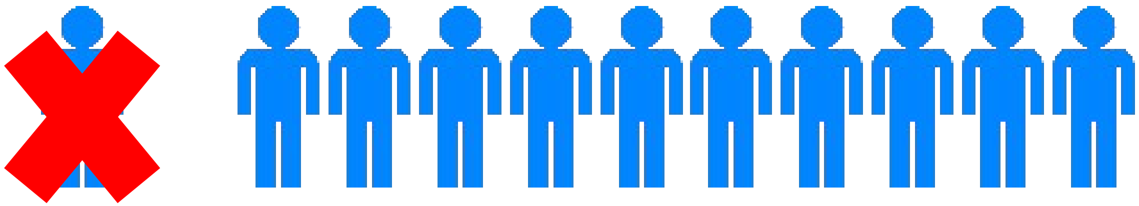
- Discrimination on positive images
- Associate the opposite class examples (VPP)



With or without help?

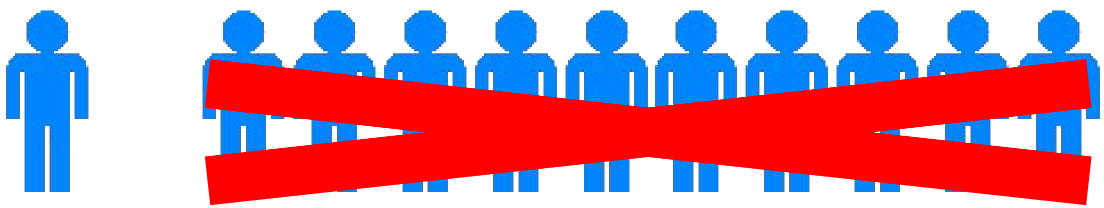
→ Which images have positive labels?

- No trust in the participant => rely on the committee (**committee validation**)



All the participants involved so far

- More trust in the participant (**user-centered**)



Context

Existing works

1

2

- Methods
- ***Results***

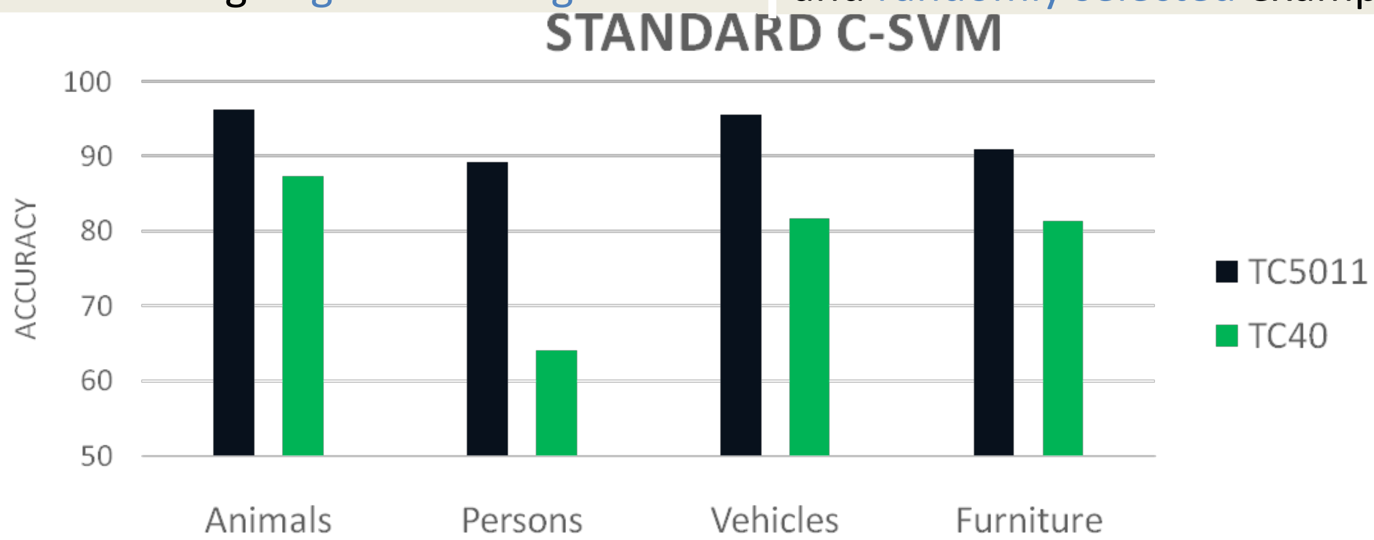
Conclusion

- Baseline C-SVM
- powerSVM
- P-SVM in Committee Validation
- P-SVM in User-Centred
- Food application

Baseline C-SVM (accuracy)

→Is it possible to classify images according to general categories?

→100-time smaller training set and randomly selected examples

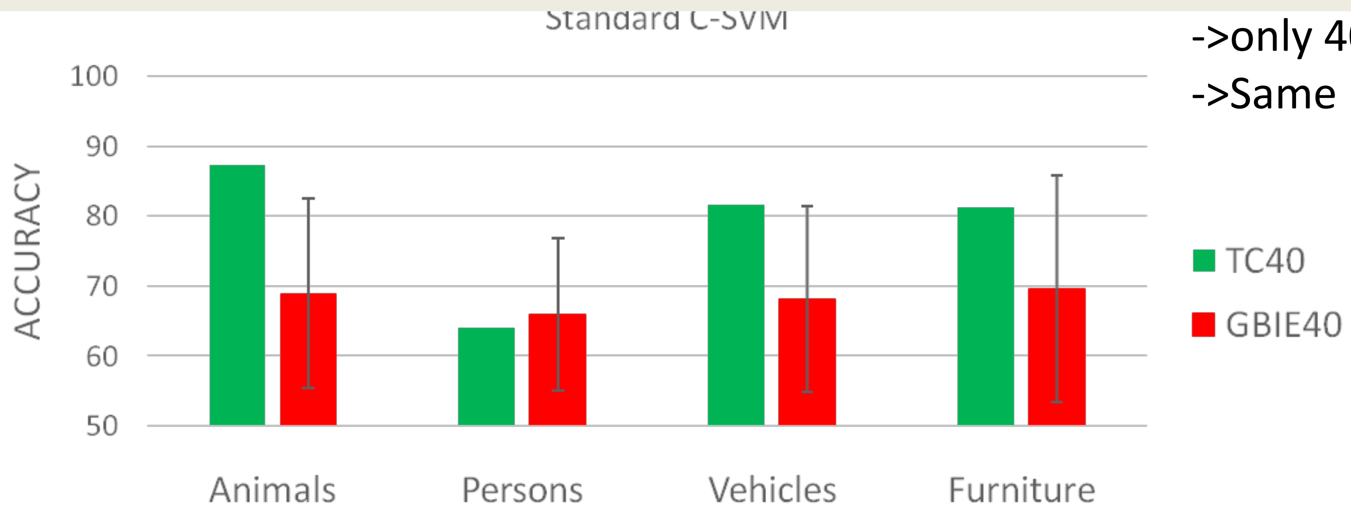


Results	Impacts	Perspectives
• General categories • 100-time smaller training set with no optimally chosen images	• Learn general categories • Learn from smaller training sets is possible	• Optimize the selection of images

Baseline C-SVM (accuracy)

→100-time smaller training set and randomly selected examples

+ GBIE labels



->only 40 images

->Same images

Results	Impacts	Actions

- Noisy labels
  - 10-15% loss

- C-SVM is not robust to label noise

- Take label uncertainty into account

Context

Existing works

1

2

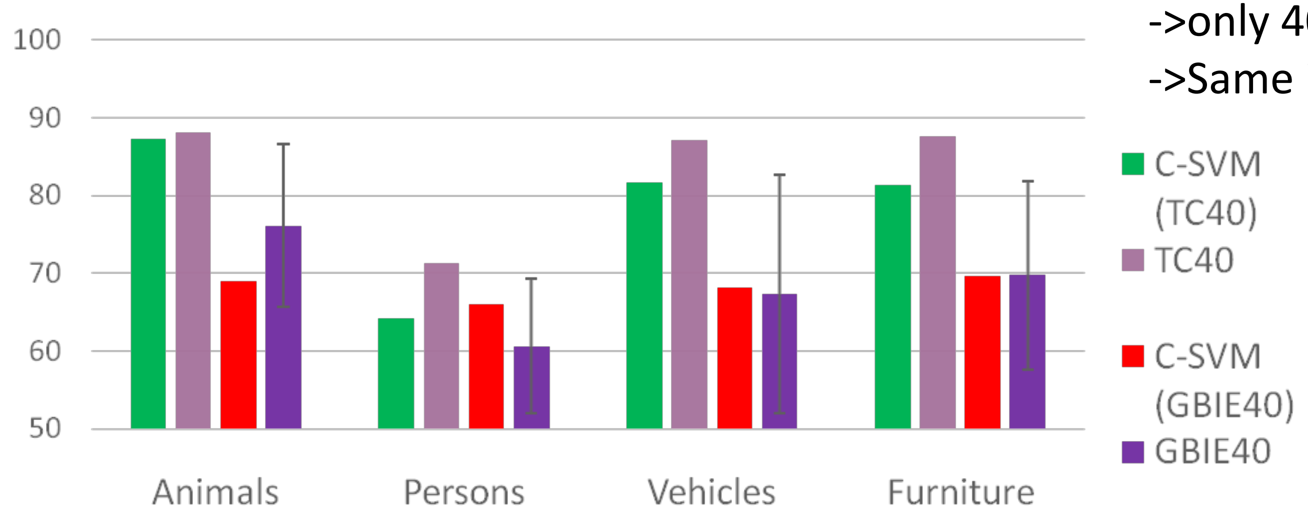
- Methods
- ***Results***

Conclusion

powerSVM: Robust?

→100-time smaller training set and randomly selected examples
+ GBIE labels

->only 40 images
->Same images



Results	Impacts	Actions

Results
- Improvement with TC40
- Not satisfactory with GBIE40

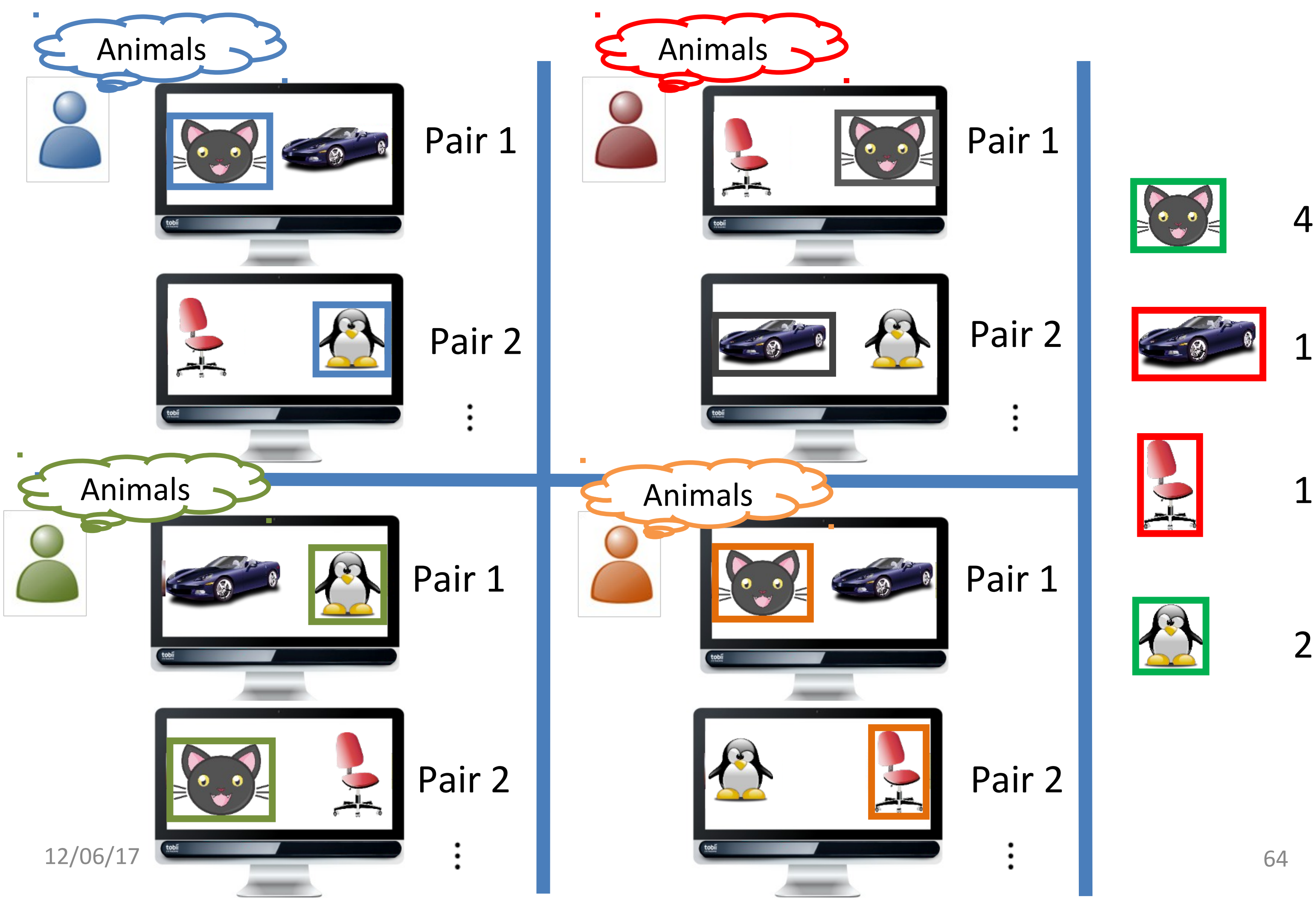
Impacts
- PowerSVM is not robust to label noise

Actions
- Find another method to handle label noise **-> P-SVM**

Committee Validation

- Context
- Existing works
- 1
- 2
  - Methods
  - ***Results***
- Conclusion

- Among all the images
  - Select half of them that have the highest number of positive labels



Context

Existing works

1

2

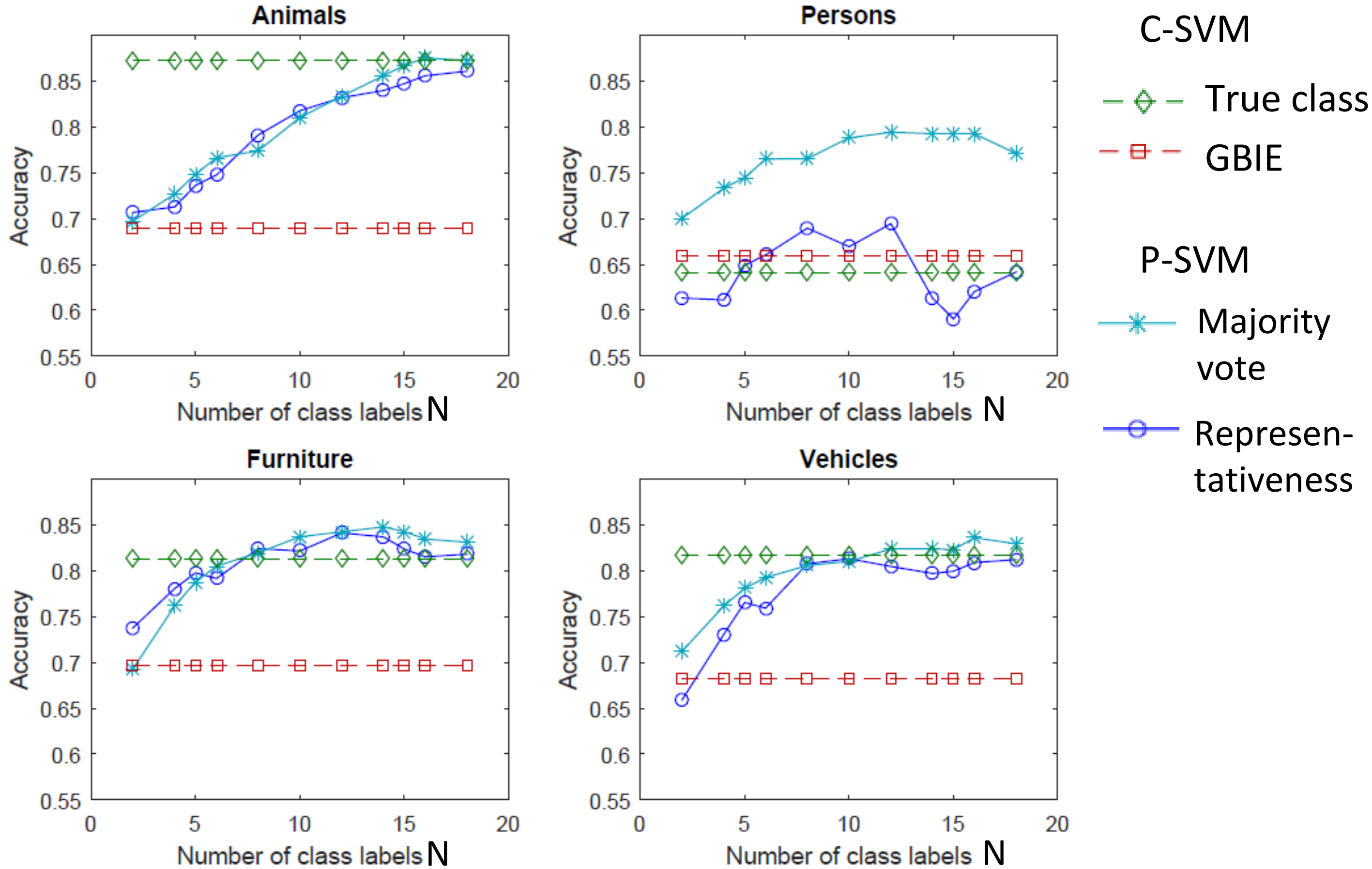
- Methods
- ***Results***

Conclusion

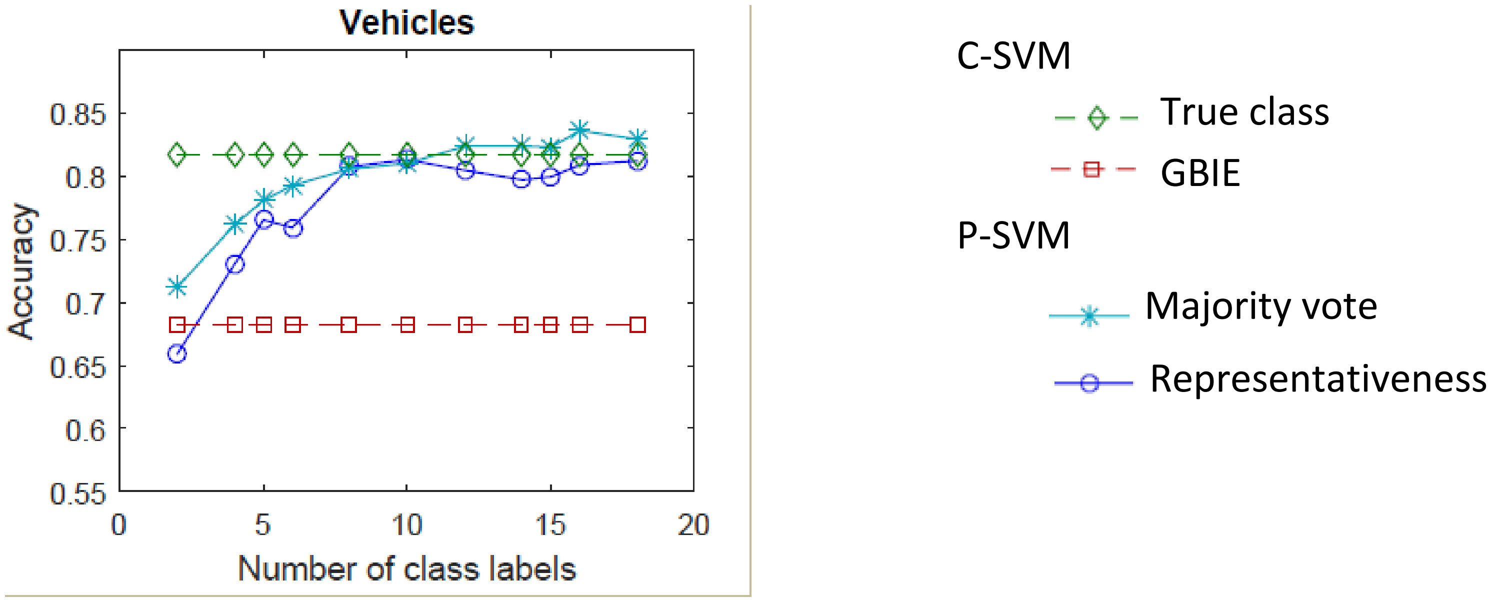
Accuracy

Function of the number of reliable labels N

Only 40 images



Accuracy



Results	Impacts	Perspectives

Results

- Both criteria get **similar results**
- With **half of the labels** considered as reliable

Impacts

- **Committee Validation context**:

GBIE with no fine-tuning is enough to annotate the training set

Perspectives

- Integration into an interactive search engine

Context

Existing works

1

2

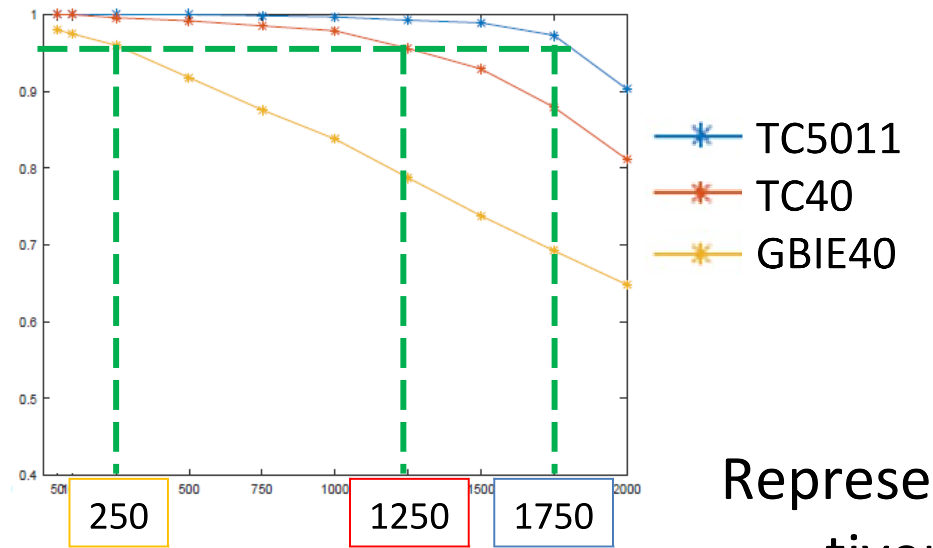
- Methods
- ***Results***

Conclusion

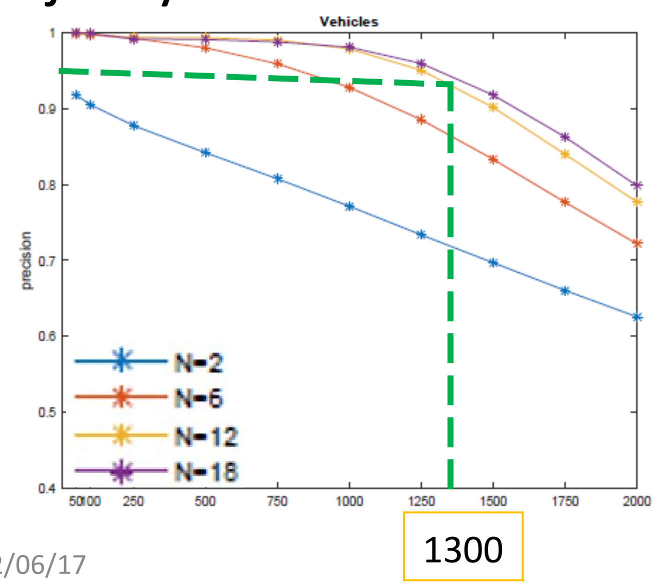
Precision@k

<u>Target: vehicles</u>

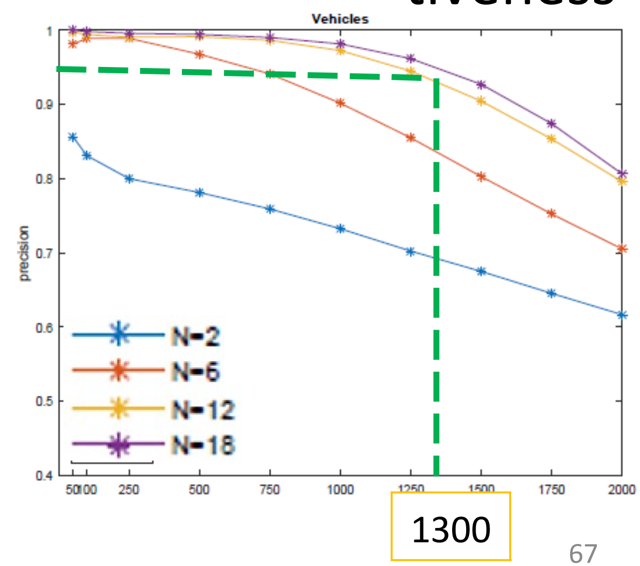
C-SVM



Majority vote

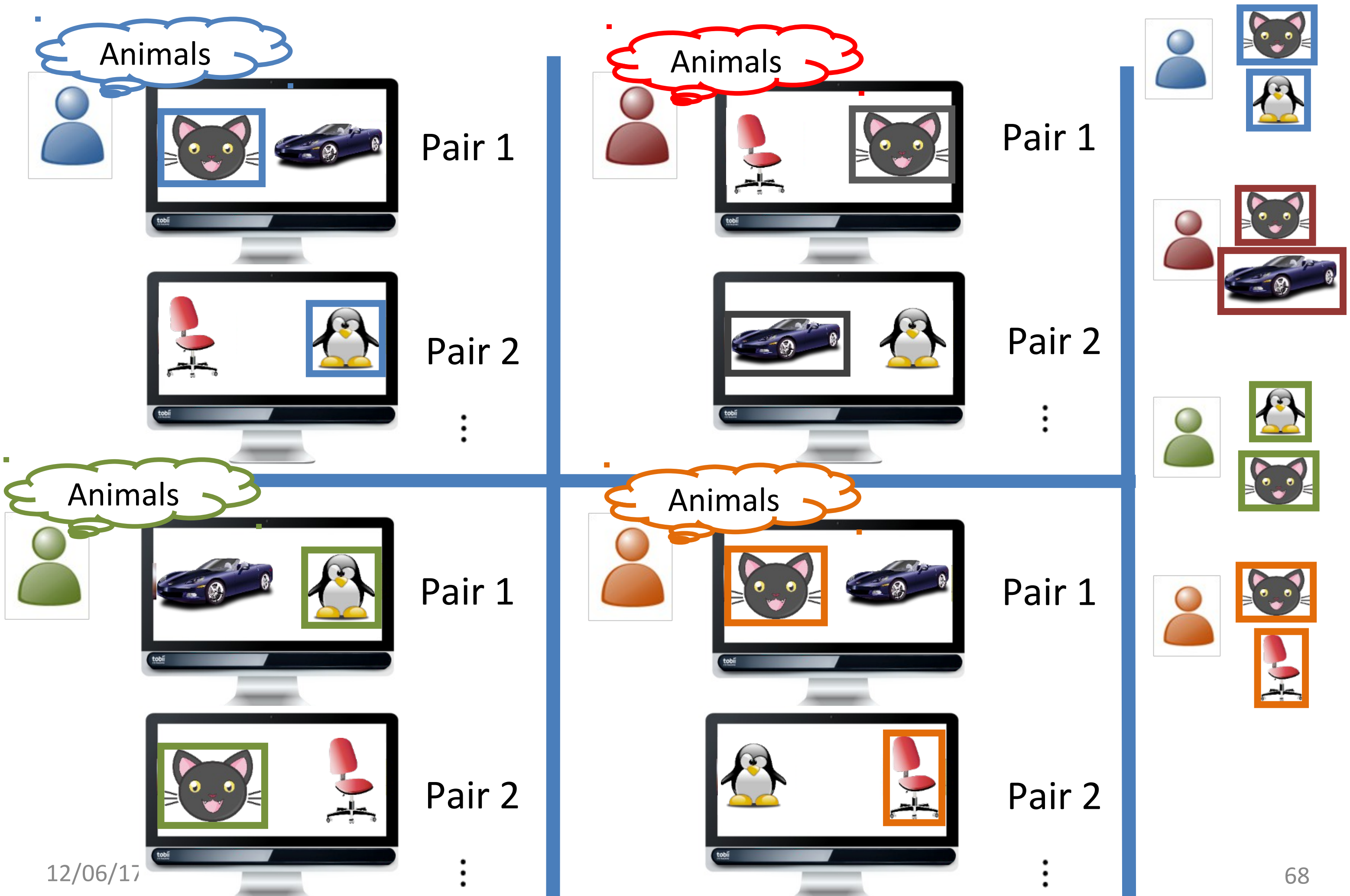


Representa-tiveness



User Centred

- Among all the images
  - Select the images that the participant has annotated as positive



Context

Existing works

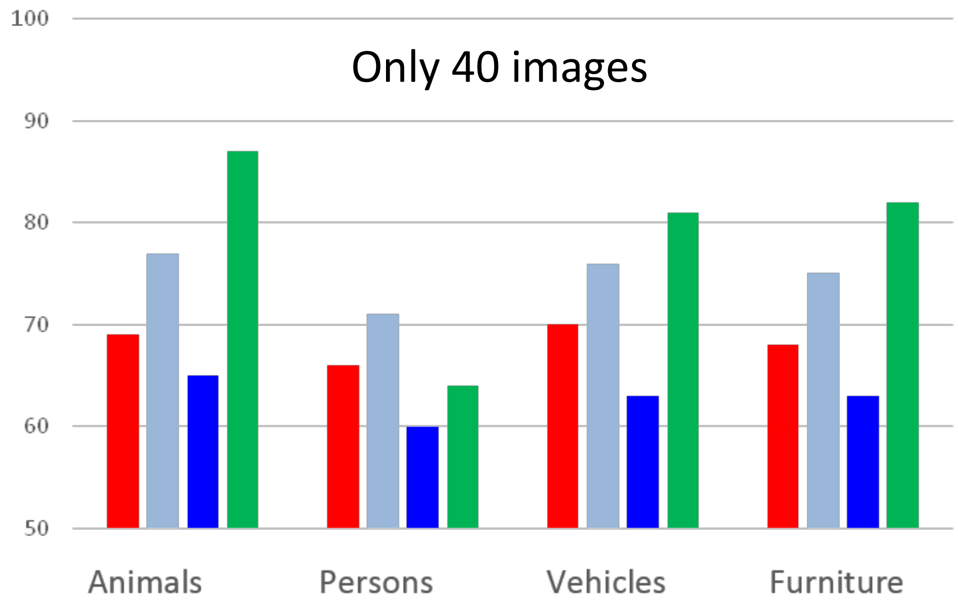
1

2

- Methods
- ***Results***

Conclusion

Accuracy



C-SVM

- GBIE40
- TC40

P-SVM

- Majority score (half of the participants consider the image as positive)
- Representativeness (score higher than 0.5)

Results

- Majority score: improvement
- Representative-ness: worse

Impacts

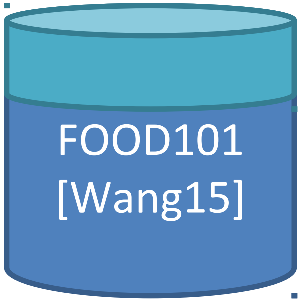
- Full user-centred approach is not possible yet

Perspectives

- Find another criterion of label discrimination without relying on users' vote

Data (food classification)

Experiment F1



FOOD 101	Training	Test
Beef carpaccio	671	224
Beet salad	664	222
Cannoli	689	230
Ice cream	694	232
Total	2718	908

Context

Existing works

1

2

- Methods
- ***Results***

Conclusion

Food Classification

C-SVM

	Beef carpaccio	Beet salad	Cannoli	Ice Cream
Original	70.0%	77.9%	77.9%	76.2%
TC40	54.6%	62.8%	46.7%	60.0%
GBIE40	51.6%	50.7%	49.3%	54.0%

Results

- Small training sets with true class labels (2): barely reach 60%

Impacts

- Small training sets with noisy labels (30%) (3)
- => Not better than random classification

Actions

- Optimize the selection of images

Extensions

Context

Existing works

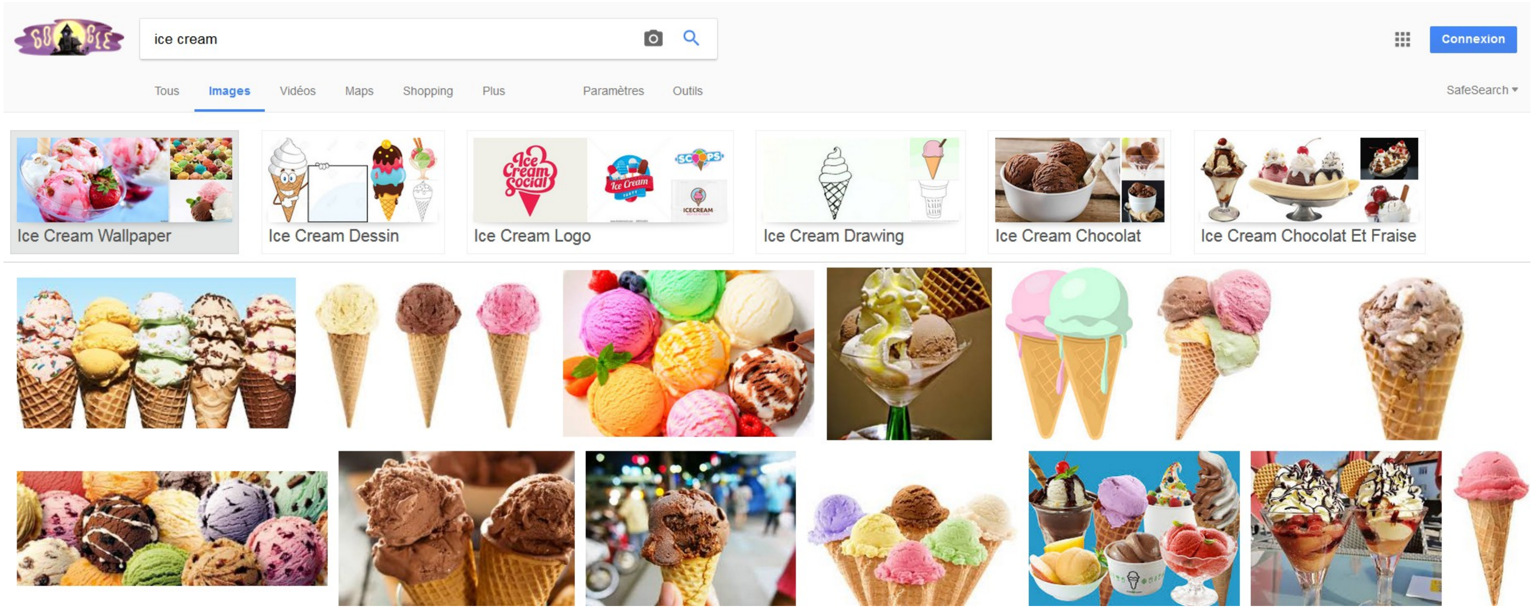
1

2

- Methods
- ***Results***

Conclusion

Extensions



	VPP	Google
• What are you looking for	60%	
• Get original ideas	60%	
• Discover a new concept		76%

Context

Existing works

1

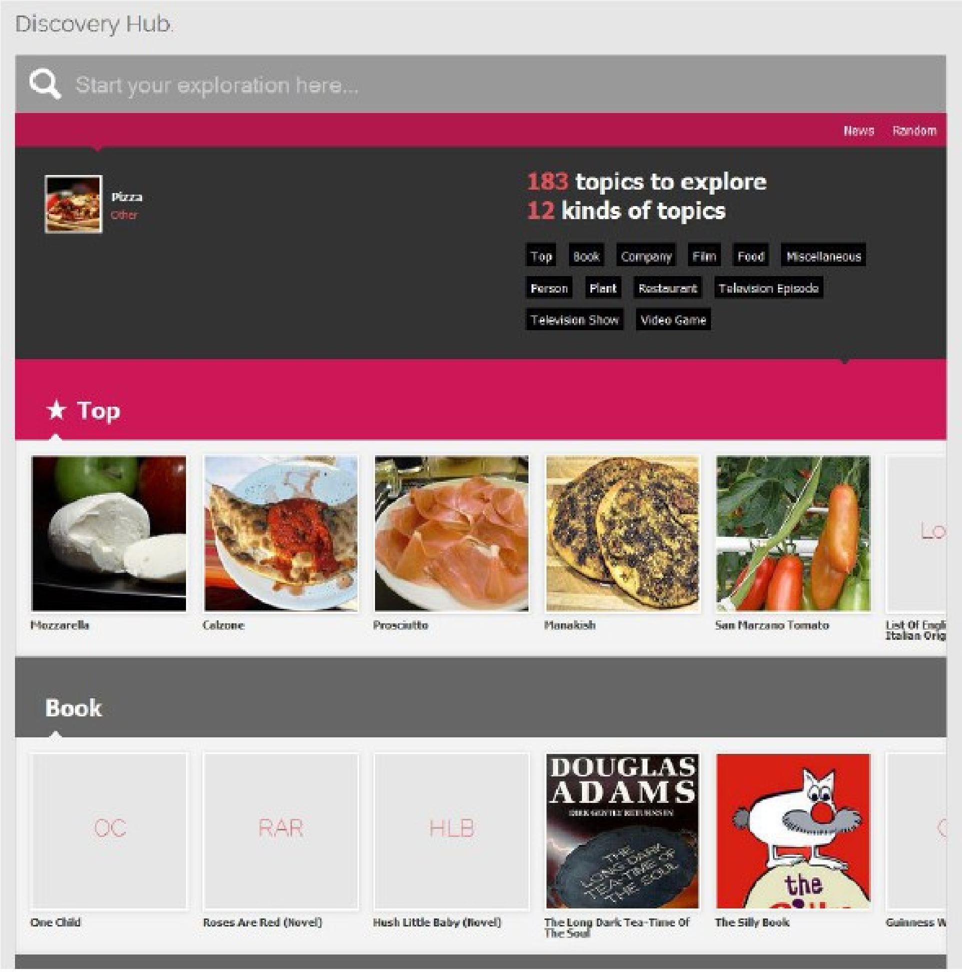
2

- Methods
- ***Results***

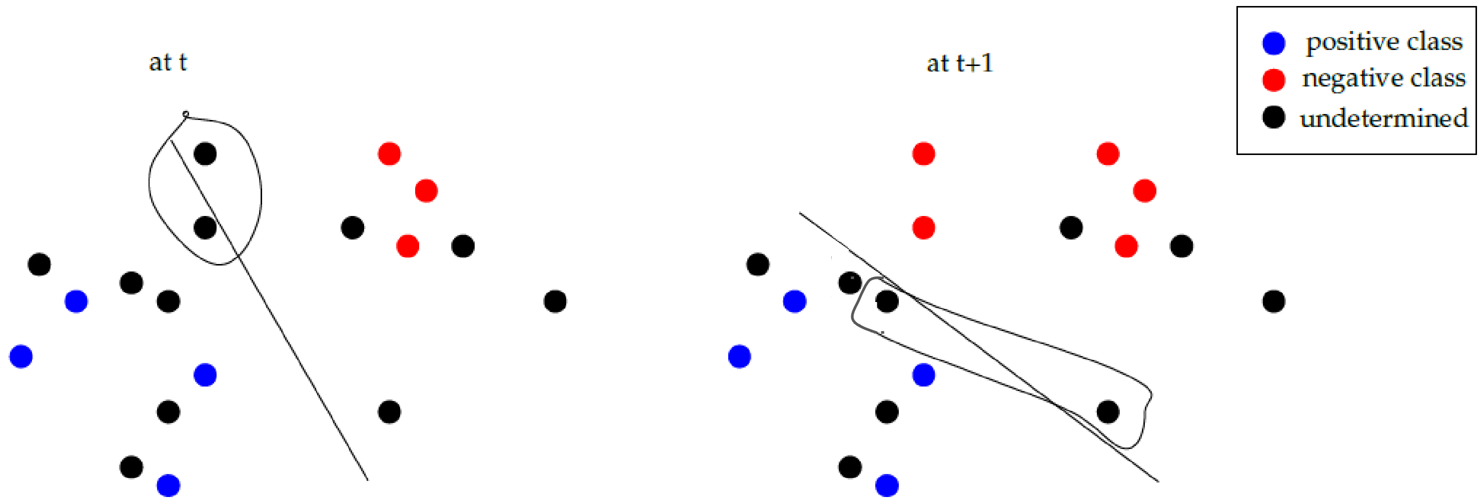
Conclusion

Extensions

Exploratory search



Optimal selection of images

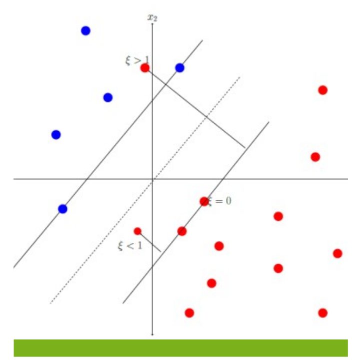


C-SVM accuracy

	Animals		Persons		Vehicles		Furniture	
TC5011	96.3 %	-	89.3 %	-	95.5 %	-	90.9 %	-
TC40	87.3 %	-	64.1 %	-	81.7 %	-	81.3 %	-
AL40	91.6 %	7%	74.9 %	7.1%	86.6 %	7%	90.5 %	7%

Summary and Future works

Context | Existing works | Annotation | Classification | Conclusion



Context

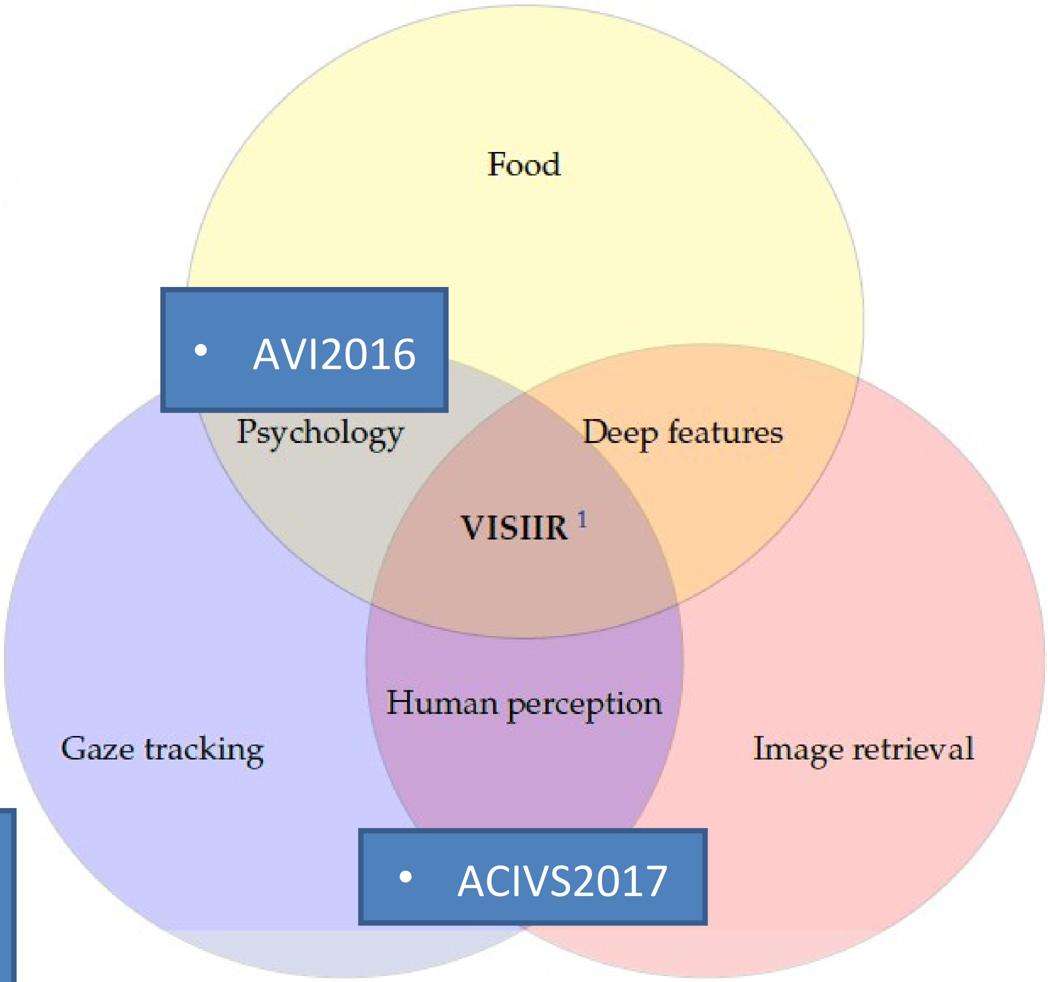
Existing works

1

2

Conclusion

Contributions



- AVI2016

- ICIP2015
- Publicly available gaze data

- ACIVS2017

Summary

GBIE annotations

- Limits the burden of manual annotation
- User independent
- Category independent
- Real-time decision

Classification purpose

- **powerSVM**: representativeness score
- **P-SVM**: handling uncertain labels
- Committee Validation

Extensions

- User applications in explatory search
- Active learning compliance

Future works

GBIE annotations

- Study temporal features
- Robustness to low resolution webcams
- More complex interfaces [Hajimirza12]

User-centred context

- Find a criterion that is related to image features [Tudor16]

Active learning

- Select iteratively the images to display in an interactive learning flavour
- What do you want + What is it? [Wang2017]

Souad Chaabouni

This work is partly funded by French National Agency for Research, VISIIR project, under contract number ANR-13-CORD-0009.

THANK YOU!

Do you have any question?

UNIVERSITÉ CÔTE D'AZUR

Visual Seek for Interactive Image Retrieval