Using the Forest to see the Trees: A computational model relating features, objects and scenes

Antonio Torralba
CSAIL-MIT

Joint work with
Aude Oliva, Kevin Murphy, William Freeman
Monica Castelhano, John Henderson
From objects to scenes

SceneType $\in \{\text{street, office, ...}\}$

Object localization

Local features

Image

Riesenhuber & Poggio (99); Vidal-Naquet & Ullman (03); Serre & Poggio, (05); Agarwal & Roth, (02); Moghaddam, Pentland (97), Turk, Pentland (91), Vidal-Naquet, Ullman, (03) Heisele, et al, (01), Agarwal & Roth, (02), Kremp, Geman, Amit (02), Dorko, Schmid, (03) Fergus, Perona, Zisserman (03), Fei Fei, Fergus, Perona, (03), Schneideman, Kanade (00), Lowe (99)
From scenes to objects

SceneType ∈ \{street, office, …\}

Object localization

Global gist features

Local features

Image

Object localization

SceneType ∈ \{street, office, …\}
From scenes to objects

SceneType ∈ {street, office, …} → S

Object localization

O₁ O₁ O₁ O₁ O₂ O₂ O₂ O₂

Local features

L L L L L

Image

Global gist features

G
The context challenge

What do you think are the hidden objects?

Biederman et al 82; Bar & Ullman 93; Palmer, 75;
The context challenge
What do you think are the hidden objects?

Answering this question does not require knowing how the objects look like. It is all about context.

Chance ~ 1/30000
From scenes to objects

SceneType $\in \{\text{street, office, …}\}$

Local features

Global gist features

Image
Scene categorization

Office

Corridor

Street

Oliva & Torralba, IJCV’01; Torralba, Murphy, Freeman, Mark, CVPR 03.
Place identification

Office 610

Office 615

Draper street

59 other places…

Scenes are categories, places are instances
Supervised learning

\[
\begin{align*}
&\rightarrow \{ V_g, \text{Office} \} \\
&\rightarrow \{ V_g, \text{Office} \} \\
&\rightarrow \{ V_g, \text{Corridor} \} \\
&\rightarrow \{ V_g, \text{Street} \} \\
\vdots
\end{align*}
\]

Classifier
Supervised learning

Which feature vector for a whole image?
Global features (gist)

First, we propose a set of features that do not encode specific object information

Oliva & Torralba, IJCV’01; Torralba, Murphy, Freeman, Mark, CVPR 03.
Global features (gist)

First, we propose a set of features that do not encode specific object information

Oliva & Torralba, IJCV’01; Torralba, Murphy, Freeman, Mark, CVPR 03.
Example visual gists

Global features (I) ~ global features (I’)

Learning to recognize places

We use annotated sequences for training

- Hidden states = location (63 values)
- Observations = $v^G_t$ (80 dimensions)
- Transition matrix encodes topology of environment
- Observation model is a mixture of Gaussians centered on prototypes (100 views per place)
Wearable test-bed v1

Kevin Murphy
Wearable test-bed v2
Place/scene recognition demo

t=930, truth = 400-fl6-visionArea1
From scenes to objects

SceneType ∈ \{street, office, ...\}

Object localization

Local features

Global gist features

Image
Global scene features predicts object location

New image

Image regions likely to contain the target

$v_g$
Global scene features predicts object location

Training set (cars)

$\rightarrow \{ V_g^1, X^1 \}$

$\rightarrow \{ V_g^2, X^2 \}$

$\rightarrow \{ V_g^3, X^3 \}$

$\rightarrow \{ V_g^4, X^4 \}$

$\vdots$

The goal of the training is to learn the association between the location of the target and the global scene features.
Global scene features predicts object location

Results for predicting the vertical location of people

Results for predicting the horizontal location of people
The layered structure of scenes

In a display with multiple targets present, the location of one target constrains the ‘y’ coordinate of the remaining targets, but not the ‘x’ coordinate.
Global scene features predicts object location

Stronger contextual constraints can be obtained using other objects.
Attentional guidance

Saliency models: Koch & Ullman, 85; Wolfe 94; Itti, Koch, Niebur, 98; Rosenholtz, 99
Attentional guidance

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003
Attentional guidance

Local features

Global features

Saliency

Object model

Scene prior

TASK
Comparison regions of interest

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003
Comparison regions of interest

Saliency predictions

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003
Comparison regions of interest

Saliency predictions

Saliency and Global scene priors

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003
Comparison regions of interest

Dots correspond to fixations 1-4

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003
Comparison regions of interest

Saliency predictions

Dots correspond to fixations 1-4

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003
Results

% of fixations inside the region

Scenes without people

Scenes with people

Chance level: 33 %

- Saliency Region
- Contextual Region
Task modulation

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003
Task modulation

Saliency predictions

Saliency and Global scene priors

Mug search

Painting search
Discussion

• From the computational perspective, scene context can be derived from global image properties and predict where objects are most likely to be.

• Scene context considerably improves predictions of fixation locations. A complete model of attention guidance in natural scenes requires both saliency and contextual pathways.